CIVIL ENGINEERING STUDIES
Illinois Center for Transportation Series No. 20-012
UILU-ENG-2020-2012
ISSN: 0197-9191

Dynamic Travel Time Estimation for Northeast Illinois Expressways

Prepared By
Abolfazl Mohammadian, PhD
Homa Taghipour
Amir Bahador Parsa
University of Illinois at Chicago

Research Report No. FHWA-ICT-20-2007

A report of the findings of

ICT PROJECT R27-177
Dynamic Travel Time Estimation for Northeast Illinois Expressways

https://doi.org/10.36501/0197-9191/20-012

Illinois Center for Transportation

June 2020

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FHWA-ICT-20-007	N/A	N/A

4. Title and Subtitle	5. Report Date
Dynamic Travel Time Estimation for Northeast Illinois Expressways	June 2020
	6. Performing Organization Code N/A

7. Authors	8. Performing Organization Report No.
Abolfazl Mohammadian (https://orcid.org/0000-0003-3595-3664) Homa Taghipour (https://orcid.org/0000-0001-8450-5775) Amir Bahador Parsa (https://orcid.org/0000-0001-8670-9894)	ICT 20-012 UILU-ENG-2020-012

9. Performing Organization Name and Address	10. Work Unit No.
Illinois Center for Transportation Department of Civil and Environmental Engineering University of Illinois at Urbana-Champaign 205 North Mathews Avenue, MC-250 Urbana, IL 61801	N/A **11. Contract or Grant No.** R27-177

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Illinois Department of Transportation (SPR) Bureau of Research 126 East Ash Street Springfield, IL 62704	Final Report 9/1/17–6/30/20 **14. Sponsoring Agency Code**

15. Supplementary Notes

Conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration.
https://doi.org/10.36501/0197-9191/20-012

16. Abstract

Having access to accurate travel time is critical for both highway network users and traffic operators. Travel time that is currently reported for most highways is estimated by employing naïve methods that use limited sources of data. This might result in inaccurate travel time prediction and could impose difficulties on travelers. The purpose of this report is to develop an enhanced travel time prediction model using multiple data sources, including loop detectors, probe vehicles, weather condition, geometry, roadway incidents, roadwork, special events, and sun glare. Different models are trained accordingly based on machine learning techniques to predict travel time 5 min, 10 min, and even 60 min ahead. A comparison of techniques showed that 15 min or shorter prediction horizons are more accurate when applying the random forest model, although the prediction accuracy of longer prediction horizons is still acceptable. An algorithm is proposed for dynamic prediction of travel time in which the travel time of each highway corridor is calculated by adding the predicted travel time of each link of the corridor. The proposed dynamic approach is tested and evaluated on highways and showed a significant improvement in the accuracy of predicted travel time in comparison to the snapshot travel time prediction approach. Traffic-related variables, especially occupancy, are found to be effective in short-term travel time prediction using loop-detector data. This suggests that among traffic variables collected by loop detectors, occupancy can capture traffic condition better than other variables. Fusion of several data sources, however, increases prediction accuracy of the models.

17. Key Words	18. Distribution Statement
Travel Time, Dynamic Prediction, Highways, Loop Detector, Probe Vehicle, Weather Condition, Machine Learning	No restrictions. This document is available through the National Technical Information Service, Springfield, VA 22161.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	48	N/A

ACKNOWLEDGMENT, DISCLAIMER, MANUFACTURERS' NAMES

This publication is based on the results of **ICT-R27-177: Dynamic Travel Time Estimation for Northeast Illinois Expressways**. ICT-R27-177 was conducted in cooperation with the Illinois Center for Transportation; the Illinois Department of Transportation; and the U.S. Department of Transportation, Federal Highway Administration.

Members of the Technical Review Panel (TRP) were the following:

- Kevin Price, TRP Chair, Illinois Department of Transportation
- Peter Stresino, Illinois Department of Transportation
- Megan Swanson, Illinois Department of Transportation
- Jeff Galas, Illinois Department of Transportation

EXECUTIVE SUMMARY

Travel time is an important component of advanced traffic-management systems, and it plays a pivotal role for both road users and traffic operators. Road users need to have access to reliable travel time information to make trip decisions, such as departure time, travel mode, or route choice before and during trips. To reduce congestion as well as increase efficiency and safety of traffic in road networks, accurate travel time information is required for traffic operators and transportation agencies to better manage and control traffic. Traffic congestion, especially in large cities in the United States, has serious health and financial consequences. Congestion costs each American on average 97 hours, or $1,348 a year (INRIX 2019). For Chicago, these numbers are even worse: 138 hours, or $1,920 a year. Providing accurate travel time information could help road users avoid congested roads, resulting in a better distribution of traffic on the road network.

On most highways, Dynamic Message Signs report the travel time, which is estimated by employing naïve methods that use limited sources of data. While this travel time estimation utilizes only limited data sources, its prediction accuracy might be not acceptable because it essentially considers snapshots in time. Therefore, it is important to develop a method that utilizes available data sources and appropriate methodologies to predict travel time accurately yet immediately. The goal of this project was to investigate the effect of different factors, including traffic-related variables, road geometry, weather condition, roadway incidents, roadwork, special events, and sun glare, on travel time prediction of highways. This was conducted by combining multiple data sources, employing data-driven techniques, considering the dynamic nature of travel time, and developing an enhanced model to predict travel time accurately.

This project was conducted in three steps. The first step was to prepare the main dataset. Datasets were extracted from different sources, including loop detectors, probe vehicles, weather condition, roadway incidents, roadwork, special events, and sun glare. Then, the datasets were analyzed, cleaned, and combined to prepare the main dataset for modeling. In the second step, different data-driven methods were investigated to estimate and predict travel time using the combined dataset. The best model was selected for the next step. The third step was to develop a dynamic approach for short-term travel time prediction on highway corridors.

A comparison of different techniques and variable sets showed that 15 min or shorter prediction horizons were more accurate when applying the random forest model, although the prediction accuracy of longer prediction horizons was still acceptable. We proposed an algorithm for dynamic prediction of travel time. In the algorithm, travel time of a corridor was calculated by adding the predicted travel time of each link of the corridor. Travel time was then predicted for the expected moment at which a vehicle arrives at that link. The proposed approach was tested and evaluated on highways and showed a significant improvement in the accuracy of predicted travel time in comparison to the snapshot approach.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Having sufficient information about travel time is a key component of advanced traffic-management systems, and it plays a pivotal role for both road users and traffic operators. It is difficult for road users to make decisions (e.g., departure time, travel mode, or route choice) before and during their trips (Zhang and Rice 2003; Ben-Akiva et al. 2001), unless they have access to reliable information. Having access to accurate travel time will also enable traffic operators and transportation agencies to better manage and control traffic in order to reduce congestion as well as increase efficiency and safety of traffic in road networks. Providing accurate travel time information to road users could help drivers avoid congested roads, resulting in a better distribution of traffic and a less congested road network. Travel time can be defined as the total time required for a vehicle to pass a route from one point (origin) to another (destination) with respect to the delays that might be imposed on that vehicle.

Traffic conditions of urban expressways are generally expected to follow patterns that do not vary significantly. Nonetheless, when the volume approaches road capacity, the relationship among time-dependent traffic variables becomes complex and highly nonlinear. Several external factors such as weather condition, roadway incident occurrence, special events, roadwork, and sun glare can also dramatically affect recurrent traffic patterns in different times and circumstances. For instance, Nookala (2006) showed that inclement weather conditions could increase traffic congestion and travel time. Driver behavior, reduced sight distance, and increased following distance due to slippery conditions could cause speed reduction.

On most highways, travel time reported on Dynamic Message Signs (DMS) is estimated by employing naïve methods that use limited sources of data (Huisken and van Berkum 2003). This could cause inaccurate travel time prediction and impose delays on travelers, making travel time reports on the DMS unreliable. Recurring congestion might be expected for a typical commuter. However, drivers who plan trips based on the travel times provided by the Department of Transportation might arrive early or late to their destination. Therefore, it is important to develop a method that could utilize available data sources and appropriate methodologies to predict travel time accurately yet immediately.

Studies aimed at finding travel time can be divided into two groups: estimation and prediction. Travel time estimation studies focus on calculating travel times of trips that have completed with respect to the data captured during those trips. Travel time prediction studies concentrate on using historical and real-time data to forecast the travel time for future time intervals. There are several methods that are used to predict travel time. The difference between the methodologies stems from their complexity and accuracy, which makes each methodology suitable for specific conditions. Note that travel time has dynamic behavior, especially in long highway corridors, which consist of several consecutive links. It is more accurate to predict the travel time of each corridor link dynamically with respect to travel time of links upstream, rather than to predict travel time of the whole corridor at once.

An important and effective factor on choosing a method for estimating or predicting travel time is the available data. Although there are several methods, each of which has advantages and disadvantages, no method can be used unless adequate and related sources of data are available. There are several kinds of sensors used to obtain traffic data, and these sensors can be grouped by their ability to obtain travel time directly. The two groups are link-based and point-based sensors. Link-based sensors directly measure the travel time for a route or link by different methods such as using passive probe vehicles, active test vehicles, or license-plate matching. Point-based sensors measure travel time indirectly by using devices such as loop detectors, laser detectors, and video cameras. Data collected from link-based sensors are more accurate; however, point-based sensors are more available and could collect cost-effective, real-time data.

In this study, data-driven methods with their effective, nonlinear modeling ability are selected to dynamically predict travel time. The goal of this study is to investigate factors on travel time of highways, including traffic peak hour, weather condition, roadway incidents, roadwork, special events, and sun glare. The study considered the dynamic nature of travel time, combined data sources, and employed data-driven techniques, including machine learning models. Figure 1 shows the framework of the project.

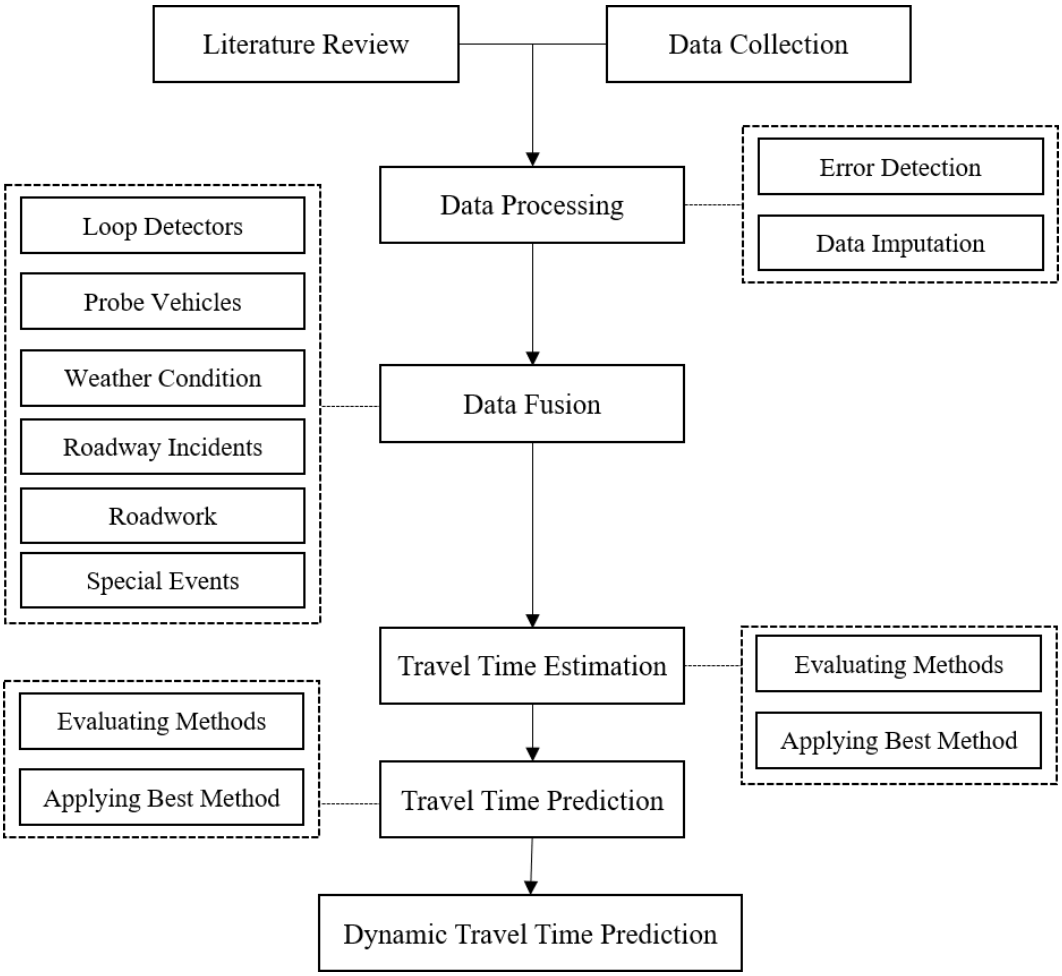


Figure 1. Project framework.

PROJECT OBJECTIVES

The main objectives of this project are defined in three parts:

1. Prepare the data and investigate methods to estimate and predict travel time.

2. Develop models for short-term travel time prediction considering the effects of all external factors.

3. Develop a dynamic approach for short-term travel time prediction.

In the first part, all datasets extracted from different sources, including loop detectors, probe vehicles, weather condition, roadway incidents, roadwork, special events, and sun glare, are analyzed, cleaned, and combined to prepare the main dataset for modeling. Different data-driven methods are investigated to estimate and predict travel time using the combined dataset. The best model is selected for use in the second and third steps. Then, a dynamic travel time prediction approach is proposed to accurately predict the experienced travel time on highway corridors.

CHAPTER 2: LITERATURE REVIEW

Numerous studies use different sources of traffic data to calculate travel time, which can be categorized into two groups: estimation and prediction. Travel time estimation studies use historical data of completed trips. Travel time prediction studies calculate the travel time for future time intervals, utilizing historical and real-time data. Various models and data sources are used for travel time calculation techniques in which trade-off between simplicity and accuracy is always a challenge. In the following sections, traffic data sources, which are mainly used for travel time calculation, are introduced. Then, different travel time prediction methods used by other researchers are presented.

DATA SOURCES

An important and effective factor of choosing a method to estimate or predict travel time is the available data. Although there are several methods, each with advantages and disadvantages, no method can be used unless adequate and related sources of data are available. There are several kinds of sensors used to obtain traffic data, which could be categorized by their technique for capturing travel time. Based on the classification done by Wu et al. (2004), there are two main groups of sensors: point-based and link-based detectors.

Point-based Detectors

Point-based detectors are sensors located at a specific point of a road that can capture traffic data at that point. Single-loop and double-loop detectors are popular examples of point-based detectors (Soriguera and Robusté 2011; Chen et al. 2011). Single-loop detectors can capture flow and occupancy variables. Double-loop detectors can also obtain point speed of vehicles. In this case, time mean speed (i.e., the average speed of all vehicles passing the detector's location) should be calculated first. Then, the time mean speed is used to calculate travel time. Note that the time mean speed cannot reflect the travel time of a corridor accurately because the point speed can vary across the route for each vehicle, especially when the road includes intersections, stop signs, on/off ramps, etc., which cause significant changes to the speed profile of vehicles (Soriguera and Robusté 2010).

Link-based Detectors

Link-based detectors can capture the space mean speed (i.e., distance traveled divided by average travel time), which could be an alternative to calculate more accurate travel time. Link-based detectors consist of two groups: floating and probe vehicles (Zheng and Van Zuylen 2013; Chen et al. 2018) and automatic vehicle identification (AVI). A common characteristic of the two types is that they are both equipped with devices such as a cell phone and Global Positioning System (GPS), so they can use these facilities to send their position, direction, and speed continually. Floating and probe vehicles capture traffic data of a specific route by passing through it. Floating and probe vehicles consist of vehicles that are employed specifically for data collection, whereas AVI comprises of passive vehicles that travel on the road network for other reasons. AVI link-based detectors consist of fixed detectors located at the beginning and end of a route, and they detect and capture traffic-related data of each vehicle at the two prespecified points of that route (Chen, Rakha, and McGhee 2013). There are several types of AVI used to calculate travel time such as automatic toll collection (El

Faouzi, Klein, and De Mouzon 2009; Fan et al. 2018), plate-matching technique (van Hinsbergen, van Lint, and van Zuylen 2009; Sánchez-Cambronero et al. 2017), as well as Bluetooth- and Wi-Fi–based detection systems (Abbott-Jard, Shah, and Bhaskar 2013; Yuan, Faghri, and Partridge 2019). Although the benefits of link-based detectors may appear to outweigh point-based detectors, data from link-based detectors are not available on all roads, whereas a relatively large number of roads are equipped with point-based detectors.

Generally, estimation and prediction are used to calculate travel time. The main issue regarding prediction is that future travel data does not exist yet, so data from past and present travels should be used in prediction methods. The two concepts are explored in further detail below.

TRAVEL TIME ESTIMATION

Travel time estimation methods can be categorized into two groups regarding how the data has been collected, as follows.

Point-based Detectors

Travel time can be estimated given data collected by single-loop and double-loop detectors. The relation between flow and occupancy can be used to estimate the travel time when single-loop detector data is available. In this case, time mean speed (i.e., the average speed of all vehicles passing the detector's location) should be calculated first. Then, the time mean speed is used to estimate the travel time. The point speed of each vehicle can be derived directly from double-loop detectors, so travel time is calculated based on the average of that speed for all vehicles. Time mean speed cannot reflect the travel time of a path segment accurately, because time mean speed of vehicles can vary across that segment. The only exception is a case in which a path segment has a steady condition regarding congestion and any other factors that are causing speed variation. It could be concluded that the more point-based detectors in a given path segment, the more accurate the travel time estimation.

Link-based Detectors

Time mean speed is not accurate enough to estimate travel time, because it can vary across the path segment. One solution to this problem is to use data from link-based detectors. If space mean speed can be driven directly, then the estimation of travel time would be more accurate. All link-based detectors could be used to capture space mean speed and, subsequently, estimate travel time.

Two points should be considered regarding point- and link-based detectors when estimating travel time. Data from link-based detectors are more likely to be noisy and have missing values in comparison to data from point-based detectors. Moreover, link-based detectors are not available on all roads, whereas a relatively large number of roads are equipped with point-based detectors (Mori et al. 2015).

Note that there are several methods and algorithms in the literature that are used to estimate travel time, using either point-based or link-based data. However, because the goal of this project is to predict travel time, exploring prediction methods and algorithms is essential. For more detailed

information, refer to the study conducted by Mori et al. (2015), which is a review of travel time estimation and prediction.

TRAVEL TIME PREDICTION

Travel time prediction is a prominent issue in transportation studies, and there are several methods that can be used to predict travel time. The difference between methodologies stems from their complexity and accuracy. Each method has unique characteristics that make it suitable for specific conditions. Models for travel time prediction can be categorized into three groups, as follows.

Naïve Models

Naïve models generally consist of very simple methods. These models have several assumptions in their nature that make them simple but less accurate. These models are generally used as the baseline for other methods or whenever a simple prediction is required.

Instantaneous predictor is a famous naïve model that assumes traffic conditions remain consistent with time (van Hinsbergen, van Lint, and van Zuylen 2009). Based on this assumption, the most up-to-date estimation of travel time is assumed to be constant and equal to travel time in the future. Expanding the prediction horizon will reduce the accuracy of this model. Historical predictor is another naïve model that has more accurate results for longer prediction horizons, unlike instantaneous predictor. This predictor uses historical travel time data to predict future travel time by regarding the similarity of traffic conditions in different periods of time (Schmitt and Jula 2007). This model works well when traffic conditions repeatedly occur in a specific path. Researchers may switch from the instantaneous model to the historical model for short- and long-term predictions, respectively. The combination of instantaneous and historical predictors is called hybrid predictor, which can be used for short- and long-term predictions. Wunderlich, Kaufman, and Smith (2000) use the weighted average of the two methods to come up with the hybrid predictor, whereas Schmitt and Jula (2007) switch between the two methods based on the required prediction horizon.

Traffic-theory-based Models

The basis of traffic-theory-based models is the relation between traffic variables such as speed, density, and occupancy. Because the prediction refers to travel time in the future, the first step in these models is to simulate traffic conditions of a network for a given time in the future. Travel time could then be predicted based on traffic condition of that network. Three kinds of simulation—macroscopic (Papageorgiou et al. 2010), microscopic (Edara et al. 2017), and mesoscopic (Taylor 2003)—are applied in various studies.

Macroscopic simulation is an approach in which fluid flow theory equations are used to simulate aggregated traffic variables (Cremer 1995). Future travel time cannot be derived from macroscopic simulation models directly, and it can be calculated based on the relation between traffic variables. Microscopic simulation takes the origin-destination matrix as an input to simulate the condition of a network under the movement of individual vehicles and the interaction between them (van Lint 2004). One of the main properties of microscopic simulation is that the travel time could be derived

directly. The last approach, which is a combination of macroscopic and microscopic simulations, is mesoscopic, which is mostly used for large networks (van Lint and van Hinsbergen 2012).

The shortcoming of traffic-theory-based models is that they require an accurate simulation in order to result in a precise prediction of travel time. However, if this requirement is met, the output would be more reliable than other prediction methods.

Data-based Models

Data-based models require a large amount of data and use statistical methods to predict travel time. The more data is available, the more accurate the results. Data-based models can be divided into two groups: parametric and nonparametric models.

Parametric Models

Parametric models have a predefined function and only the parameters of the model need to be estimated. There are several parametric models, each of which has its own structure. Famous parametric models will be introduced below.

The time-series model is a method in which the variation of past data in a system is used to forecast the future state of that system (Yang 2005a). It is assumed that at each time, the value of the series only depends on its previous values plus a random disturbance. If this relationship is linear, then the auto-regressive (AR) model is formed. However, the moving-average (MA) model assumes that the current value of the series has a linear relationship with disturbance terms. The combination of AR and MA forms the auto-regressive moving-average (ARMA) model. The auto-regressive integrated moving-average (ARIMA) model is used in cases where there is evidence of non-stationarity in the data. The "I" indicates that values of the data are replaced with the difference between their current and previous values. The ARIMA model is generally denoted as ARIMA(p,d,q) with non-negative parameters p, d, and q. The "p" is the order of the autoregressive model, "d" is the degree of differencing, and "q" is the order of the moving-average model.

In a study in Toronto, Yang (2005a) used the time-series model and travel time data to accurately predict short-term travel time. This study was conducted in a 3.7-mi corridor on a highway, and the GPS probe vehicle method was used to collect the data. This study used the ARIMA model because the observed the data was nonstationary.

Another study was conducted in the Bay Area of California on a 9-mi highway corridor (Xia, Chen, and Qian 2010). Incident data were collected from 1,613 incidents consisting of the starting time, duration, milepost, location description, and incident type. Under the incident condition, travel time should be predicted for three segments: origin to queue, queue to bottleneck, and bottleneck to destination. A weekly time series model, SARIMA, is used for traffic parameter prediction. The study showed that the proposed model can predict travel time more accurately than models that do not consider the effect of incidents.

Kalman filtering is an optimal estimation algorithm in which the state of a system is estimated by minimizing the mean-squared error of an estimation (Yang 2005b). The prediction will be made based

on a recursive procedure by using the current step's results to obtain the results of the next step. The parameters for the Kalman filtering method cannot be updated unless its trips are completed and accurate data is available. The main function of Kalman filtering is to estimate the current state of a system, but it can also predict future values for travel time or improve travel time estimation at earlier times. Prediction and correction steps are implementation steps in this method.

In a study in Minnesota, Yang (2005b) used GPS test vehicles to predict travel time as well as a recursive, discrete-time Kalman filter. The study focused on a sudden traffic surge after a special event and its effect on travel times. Yang chose a graduation ceremony at the University of Minnesota Duluth as a case study. The study found that larger errors in travel time prediction occurred when a sudden dramatic change, either increasing or decreasing, happened. Generally, Yang's predicted travel times followed the observed travel times. Yang also showed that the average error is smaller when traffic congestion lasts longer.

There are other parametric models in the literature such as linear regression (Du, Peeta, and Kim 2012) and Bayesian Network (Castillo et al. 2011), which are used to predict travel time. Please refer to Kwon, Coifman, and Bickel (2000) and Yu and Cho (2008) for further details.

Nonparametric Models

Nonparametric models are models in which the structure of the function should be defined based on the data as well as number and typology of parameters. A goal of nonparametric models is to utilize a large database to understand relationships, instead of using complex approaches, to predict travel time. Generally, to forecast the future state of a system, these methods search historical data for similar states with the current state of the system. Hence, the accuracy of these methods depends on sufficient data to detect a proper match for each state. A famous model of this group is the artificial neural network model, which has been widely used in transportation studies (Yildirimoglu and Ozbay 2012; Dharia and Adeli 2003). Other nonparametric models that are employed for travel time prediction include K-nearest neighbors (Yu et al. 2016; Zhao et al. 2018; Qiao, Haghani, and Hamedi 2013), regression trees (Nikovski et al. 2005), local regression (Simroth and Zahle 2010), support vector regression (Yildirimoglu and Ozbay 2012; Wu, Ho, and Lee 2004), and random forest (Leshem and Ritov 2007).

In recent years, neural network models are frequently used to predict the traffic state. In most studies, backpropagation neural networks (Adeli 2001) are used. To improve prediction results, other methods like counterpropagation neural networks (Dharia and Adeli 2003) are also used. Neural network models are effective in dealing with nonlinear problems and pattern recognition. They can accomplish self-learning, but they require a large dataset and have a long training process. They may also face overfitting.

In a study conducted in 2002 in the Netherlands, researchers implemented the state-space neural network model to predict freeway travel time using simulated data (van Lint, Hoogendoorn, and van Zuylen 2002). In this study, a state-space formulation was used to form a recurrent state-space neural network (SSNN) model that can deal efficiently with the spatiotemporal relationships between speed,

flow, and travel time. They tested the performance of several versions of SSNN and showed that neural network models can accurately predict experienced travel times.

Dharia and Adeli (2003) presented a method to forecast freeway travel time using counterpropagation neural network (CPNN) and simulated data. For their case study, the CPNN method and learning coefficients worked faster and with the same accuracy than the backpropagation (BP) neural network approach. They also showed that when the time steps are smaller, or the resolution of the data is higher, the error is smaller.

K-nearest neighbors is another algorithm that can be used for travel time prediction. Qiao, Haghani, and Hamedi (2013) predicted short-term travel time in a study conducted on a 1.18-mi segment of a freeway in Washington. They used Bluetooth travel time data, which could provide travel time and space mean speed directly. This study implemented four prediction models: historical average, ARIMA, Kalman filter, and K-nearest neighbors. They also proposed a modified K-nearest neighbors model, which considers traffic pattern features. The K-nearest neighbors model is implemented through the following steps:

- Step 1: Prepare a historical database with past time interval travel times.
- Step 2: Select T continuous previous intervals as a combined group, t = 1, ..., T.
- Step 3: Calculate and rank neighborhood similarity to find nearest neighbors for next interval.
- Step 4: Find a set of K-nearest neighbors.
- Step 5: Predict targeted next interval travel time by taking the average of the nearest neighbors.

The proposed model (KNN-T) uses a different formula to calculate neighborhood similarity and predict the next interval travel time. This is based on the weighted average of the next interval value of each nearest neighbor and the differential value of each nearest neighbor. In this case study, Qiao et al. showed that nonparametric models outperformed other methods. When it is possible to have access to sufficient historical data, nonparametric models can provide more accurate prediction without using complicated model calibration and computations.

DYNAMIC TRAVEL TIME PREDICTION

To predict travel time for a corridor, it is crucial to consider the dynamic nature of traffic variables. There are several studies that use various methods to increase the reliability and accuracy of travel time prediction techniques by accounting for the dynamic nature of traffic variables. Travel time prediction can be accomplished in a two-stage procedure of link-based prediction and corridor-based prediction. To investigate the traffic condition in a corridor, the snapshot and vehicle trajectory methods could be considered. This is based on the corresponding measures of snapshot travel time (i.e., adding up travel times of consecutive links at once) and experienced travel time (i.e., adding up travel times of consecutive links at the current time of each link starting from the origin) (Zeng 2011).

Because travel time has a smaller variance during peak hours, corridor-based travel time prediction is more accurate than link-based (Chien and Kuchipudi 2003). Chen and Chien (2001) used the Kalman filtering model to compare two link-based travel time prediction methods and one path-based one. They found that path-based prediction achieved the highest accuracy. In link-based prediction methods for a corridor, considering progressive addition of travel times on each link in current time period (from the origin) improves the prediction accuracy in comparison to adding travel times on all links at once.

Data-driven methods can predict travel time for future time intervals by considering the transitional nature of traffic flow without strong assumptions (Yildirimoglu and Geroliminis 2013). One data-driven method for dynamic travel time prediction is the dynamic neural network with short-term memory, which can store previous traffic information and combine it with current traffic data to predict the travel time for future time intervals (Shen 2008; Shen and Huang 2011). Yildirimoglu and Geroliminis (2013) proposed an incremental learning method to predict the experienced travel time using both historical and real-time traffic data. They clustered the dataset into homogeneous groups with similar traffic characteristics. By predicting congestion time and its evolution pattern, the expected experienced travel time for a given start time is then calculated.

Furthermore, in other studies, researchers predicted travel time dynamically using approaches such as the particle-filter approach based on trip trajectory construction (Chen and Rakha 2012) and the three-stage approach for identifying current traffic conditions. Chen, Rakha, and McGhee (2013) also searched historical data for similar traffic patterns and predicted travel time. Wang and Ma (2014) proposed a method using traffic state simulation, while Fan et al. (2018) used an adaptive travel time prognosis approach based on a random forest model.

SUMMARY AND CONCLUSION

Because loop-detector data is available for this project as the main source of data, data-based methods are employed to predict travel time. In general, most studies in this area have considered only traffic-related variables in order to develop a model for travel time prediction. However, other variables such as weather condition could have a significant impact on travel time (Qiao, Haghani, and Hamedi 2012), which is infrequently investigated in travel time prediction studies. Therefore, the goal of this project is to predict travel time for highways in the Chicago region by using a combination of data sources, including loop detectors, probe vehicles, weather condition, road network, roadway incident occurrence, roadwork, and special events.

CHAPTER 3: DATA PROCESSING AND DESCRIPTIVE ANALYSIS

Different data sources are used to predict travel time in this project. The main part of the dataset is the information captured by inductive loop detectors in the Chicago highway network. Other types of data such as incident reports, roadwork, special events, and weather condition are added to the dataset to gain more accurate results.

First, the Eisenhower Expressway in Chicago, Illinois, is selected. It is a 29.84-mi interstate highway that connects Chicago loop to the northwest suburbs of Chicago, where it meets I-90 (Figure 2). All data-cleansing methods and procedures of travel time prediction are then applied to this highway. A similar procedure is applied to other Chicago highways as well.

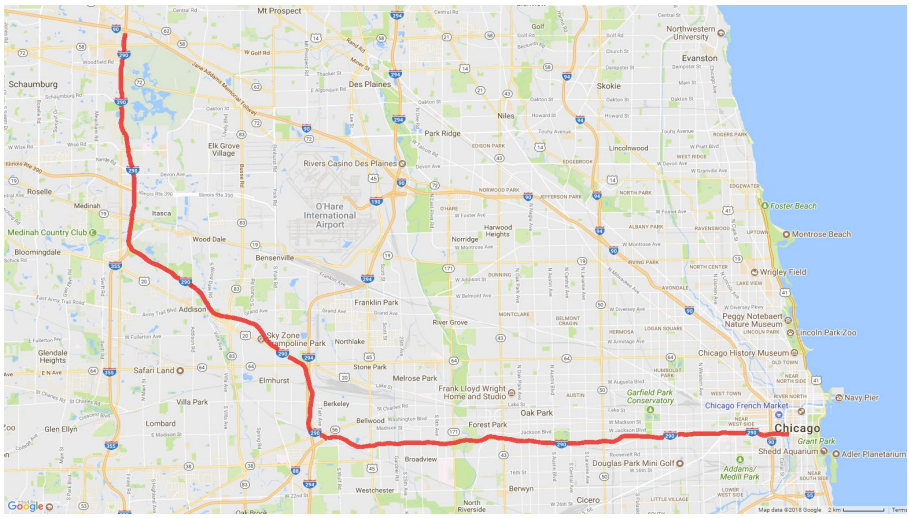


Figure 2. Eisenhower Expressway in Chicago, Illinois.

Figure 3 shows the location of loop detectors on the Eisenhower Expressway. Red and green points represent loop detectors that are located eastbound and westbound of the highway, respectively.

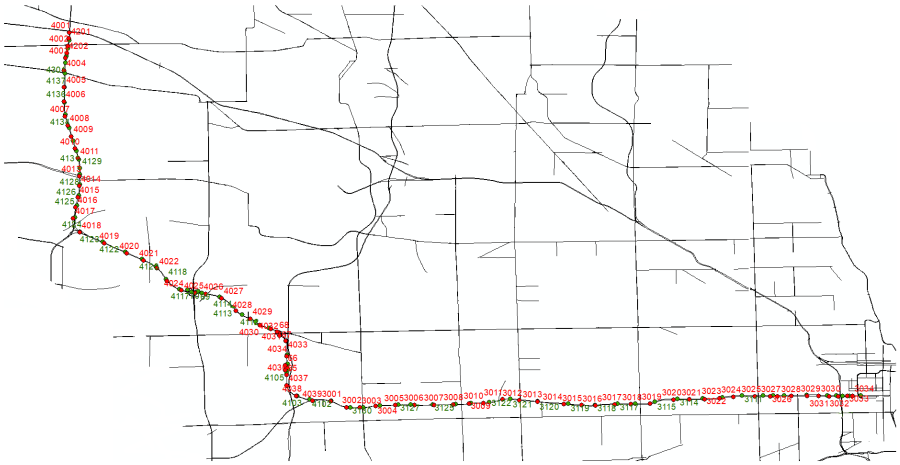


Figure 3. Loop detectors on the Eisenhower Expressway.

DATA SOURCES

Loop-Detector Data

The main dataset used in this project is the information captured by inductive loop detectors in the Chicago highway network, which is collected by the Gateway Traveler Information System and provided by the Illinois Department of Transportation (IDOT). For each loop detector per lane of the highway, the number of vehicles, occupancy, and average speed are reported every 20 sec. These raw data points contain missing and erroneous records, which might be caused by malfunctioning detectors, deteriorating pavement, or other reasons. Table 1 shows the average speed, occupancy, and flow reported in September 2017 by loop detectors in the horizontal section of the westbound Eisenhower Expressway. Note that no data is reported from the first six loop detectors.

Table 1. Average Values Reported in September 2017 by Sample Loop Detectors

Detector ID	Average Speed	Average Occupancy	Average Flow
3101	0	0	0
3102	0	0	0
3103	0	0	0
3104	0	0	0
3105	0	0	0
3106	0	0	0
3107	114.23	8.25	6.06
3108	86.03	10.76	6.17
3109	98.45	10.44	6.28
3110	108.01	10.86	6.21
3111	100.62	12.12	5.84
3112	107.59	12.26	6.44
3113	86.13	18.92	5.93
3114	86.12	15.22	6.33
3115	81.03	15.92	5.51
3116	89.51	16.64	6.06
3117	100.75	17.95	6.53
3118	80.15	18.64	6.55
3119	90.28	15.55	6.93
3120	68.63	19.30	6.71
3121	91.75	15.84	7.36
3122	87.54	14.49	8.23
3123	109.19	12.15	7.69
3124	88.24	13.69	7.56
3125	101.51	12.34	7.52
3126	90.85	12.94	7.96
3127	102.63	11.91	7.72
3128	95.33	11.74	7.66
3129	95.60	11.26	7.25
3130	105.23	10.63	6.37

Shortcomings of the raw dataset are introduced, including missing or duplicate data as well as erroneous records. By using the methods presented in the literature, it is possible to prepare the data for use in the travel time prediction procedure. Two types of single and combined thresholds are applied to the dataset as part of the data-cleaning procedure.

Single Threshold

There are several thresholds for the flow, occupancy, and speed values, which are used in similar studies. In this study, based on the characteristics of the selected highway, a value of 3,000 vehicles per hour per lane is set as the threshold for the count, which is almost equal to 17 vehicles per 20 sec per lane. If the value of count in a lane does not lie within the domain of (0, 17), it is labeled an erroneous data point. Moreover, extreme amounts for occupancy (i.e., more than 90% in 5 min) are assumed to be incorrect so that any data point with an occupancy value outside the domain (0, 90) is labeled as an erroneous data point. In addition, 80 mph is set as the upper threshold of speed, and any point with more than 80 mph or less than 0 mph speed is considered an outlier.

Combined Threshold

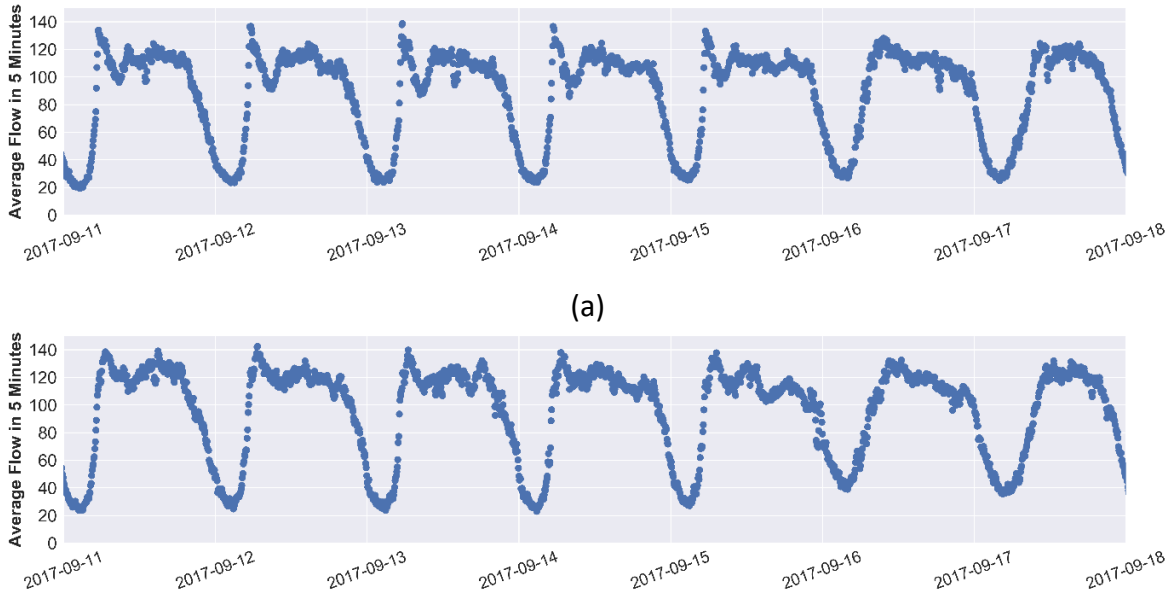
For the three variables (count, occupancy, and speed) reported by loop detectors, the first combined threshold is described as “only one zero variable out of three.” This means that if only one of the three variables is equal to zero, whereas the other two are non-zero, then that record is considered an erroneous record. The second combined threshold is “only one non-zero variable out of three.” If two of the variables are zero while the other one is non-zero, then this record is labeled an error. Finally, the last combined threshold is “all zero variables,” in which a record is considered an error if all three variables are zero. This threshold filters out the data points that could incorrectly impact the average speed of corridors.

Data Imputation

Missing and erroneous data points could negatively affect the results of prediction models. These points should either be eliminated from the dataset or imputed by new data points through one of three techniques: temporal, spatial, or historical estimation. Temporal estimation is used when a detector has a missing data point or reports a value that exceeds thresholds for a time interval. In this case, an appropriate way to impute the data point is by taking the average of values for previous and next time intervals, especially when the time intervals are short (i.e., 20 sec). Spatial estimation is used when missing or erroneous data points are imputed with the average of the values for previous and next loop detectors at the same time. This method is appropriate whenever loop detectors are close to each other or using the temporal estimation is not applicable. Historical estimation can be used whenever traffic condition is recurrent in a location in which missing or erroneous data is reported. In this case, erroneous data points are imputed using reported data points from the same location, time of day, and day of week in past records.

Figure 4 shows a sample weekly pattern of the average flow after applying the data imputation process for the links of the eastbound and westbound Eisenhower Expressway. This figure displays seven daily traffic patterns, beginning Monday, September 11, 2017, and ending Sunday, September 17, 2017. Weekdays have similar traffic patterns, although Saturday and Sunday are different. Figure 5 shows a sample daily flow pattern for the eastbound and westbound Eisenhower Expressway.

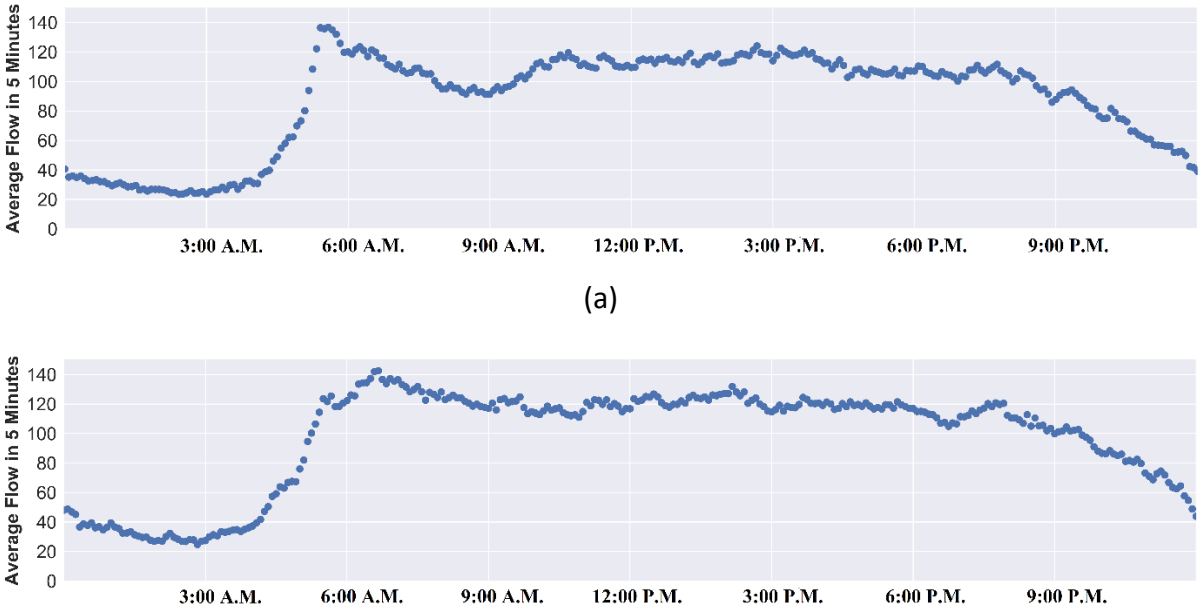
Figure 6 demonstrates the distribution of flow per 20 sec on each lane of the eastbound and westbound Eisenhower Expressway for both links.



(a)

(b)

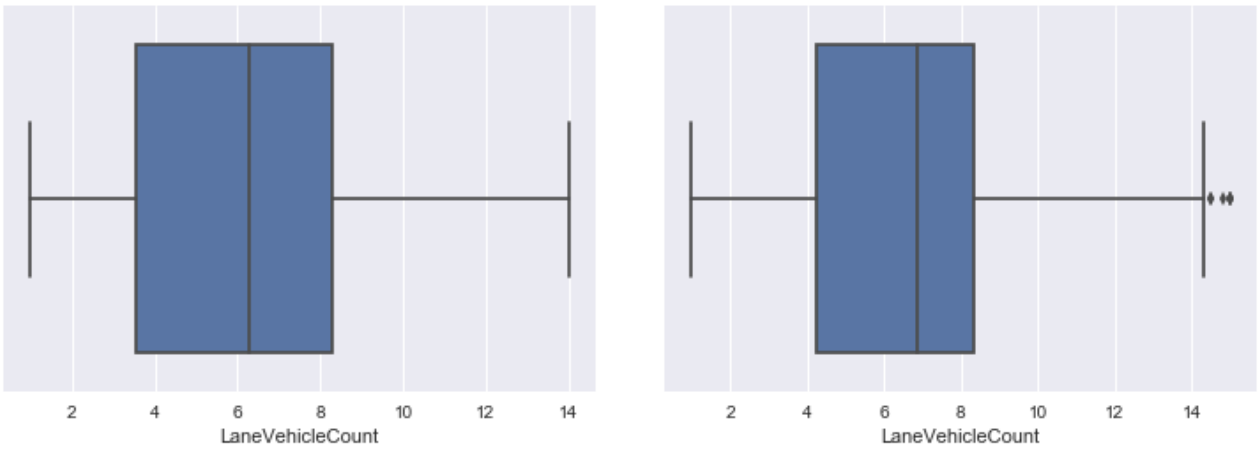
Figure 4. Weekly average flow per lane for Eisenhower Expressway links from cleaned loop-detector dataset (a) eastbound and (b) westbound (September 2017).



(a)

(b)

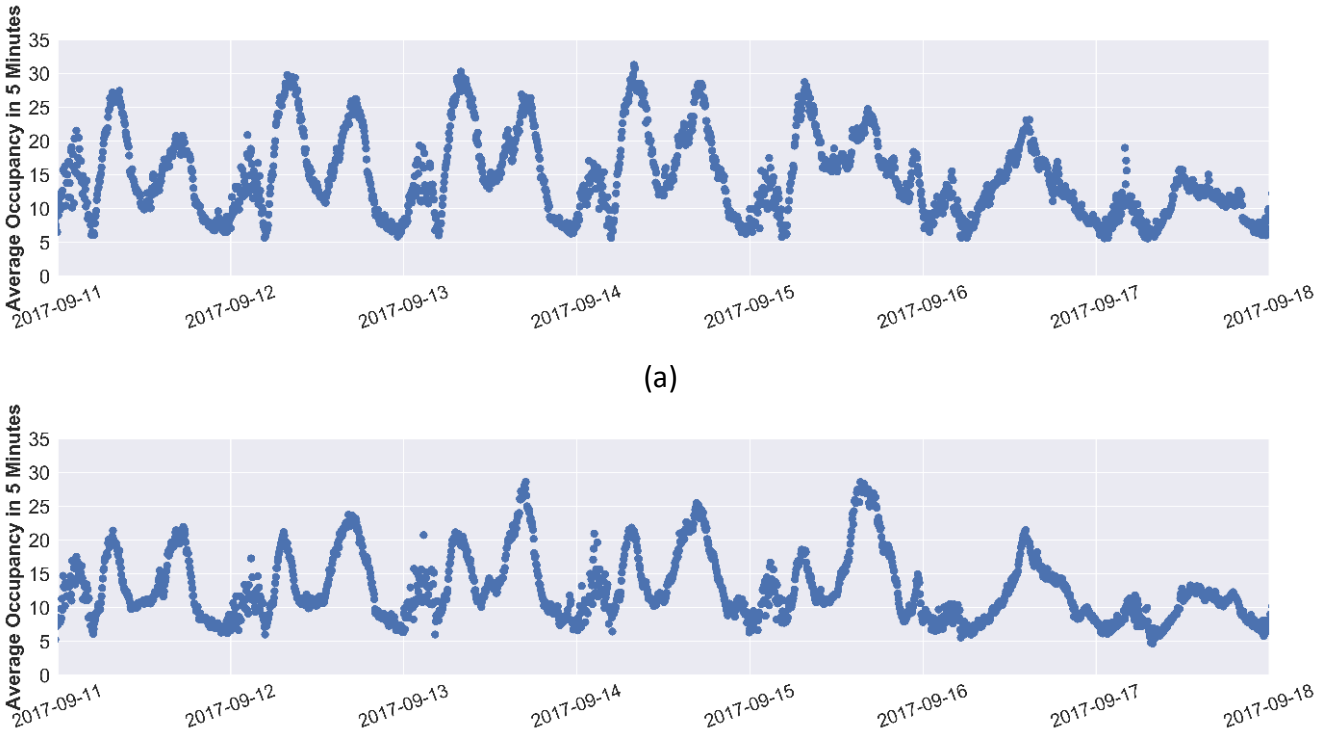
Figure 5. Daily average flow per lane for Eisenhower Expressway links from cleaned loop-detector dataset (a) eastbound and (b) westbound (September 12, 2017).



(a) (b)

Figure 6. Distribution of flow (per 20 sec per lane) for Eisenhower Expressway links from cleaned loop-detector dataset (a) eastbound and (b) westbound (September 2017).

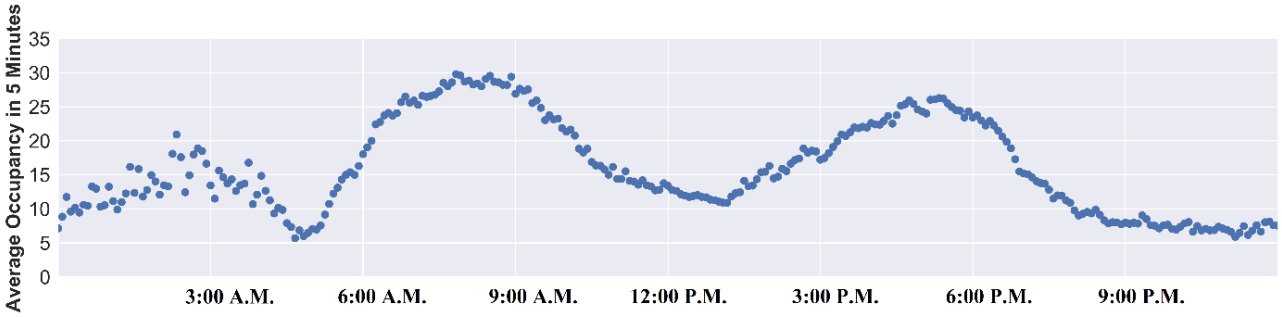
Similar plots are displayed in Figure 7, Figure 8, and Figure 9 for the average occupancy and occupancy distribution of the links of the eastbound and westbound Eisenhower Expressway.



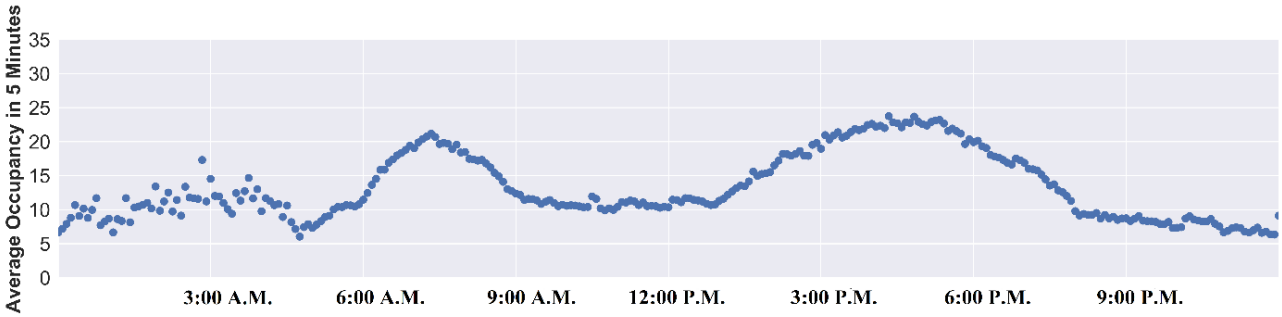
(a)

(b)

Figure 7. Weekly average occupancy of Eisenhower Expressway links from cleaned loop-detector dataset (a) eastbound and (b) westbound (September 2017).

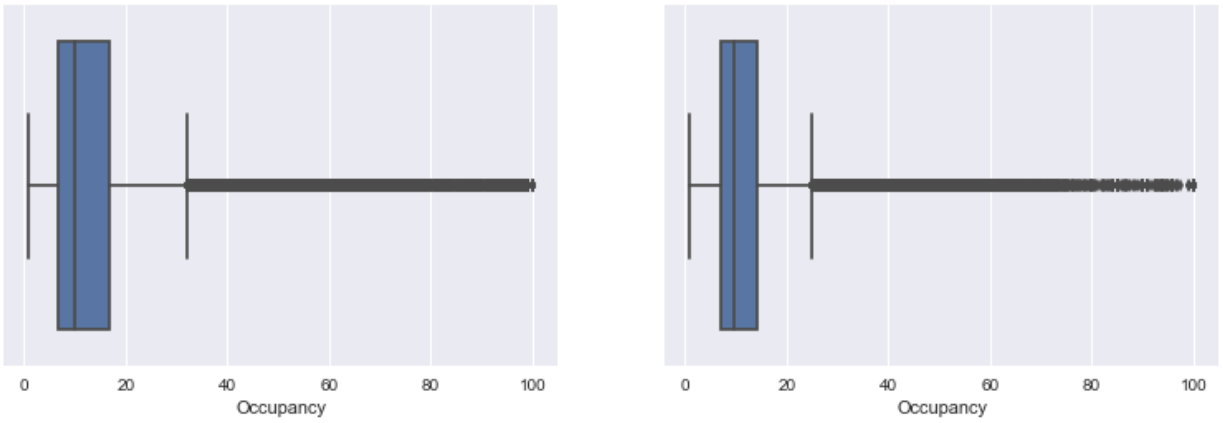


(a)



(b)

Figure 8. Daily average occupancy of Eisenhower Expressway links from cleaned loop-detector dataset (a) eastbound and (b) westbound (September 12, 2017).



(a) (b)

Figure 9. Distribution of occupancy for Eisenhower Expressway links from cleaned loop-detector dataset (a) eastbound and (b) westbound (September 2017).

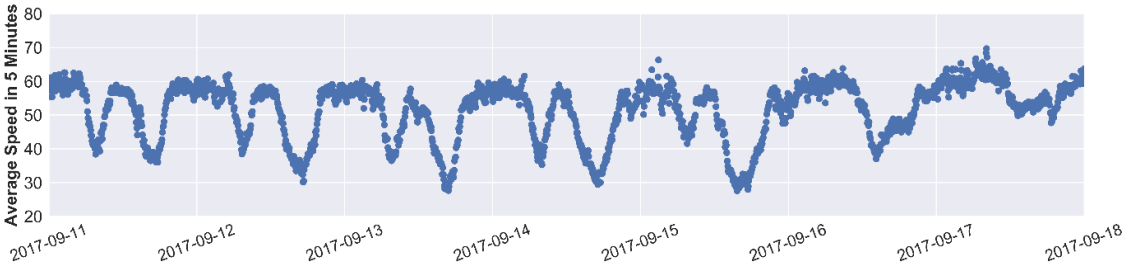
The speed values in the loop-detector dataset are time mean speeds. The time mean speed is calculated by averaging the point speeds of all vehicles passing a loop detector at that detector's location. This speed could not reflect the travel time of a corridor accurately, because the speed of vehicles could vary across that corridor. Therefore, to consider the speed values, another source of data that is collected by probe vehicles is utilized.

Probe-Vehicle Data

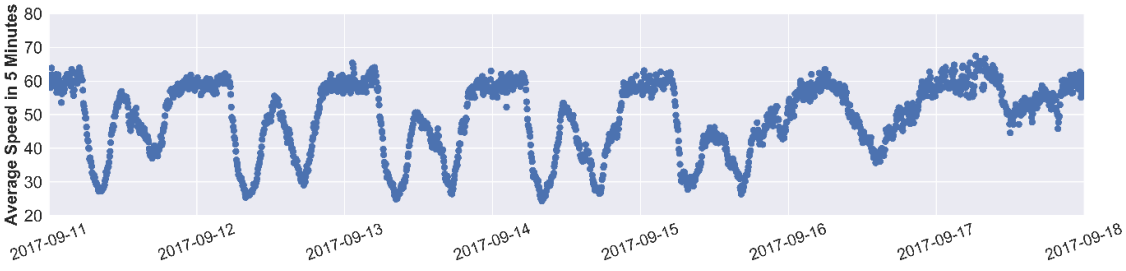
Because the speed values from the loop-detector data source are not accurate, the probe-vehicle data source is used for the speed values. This dataset is provided by the National Performance Management Research Data Set (NPMRDS) and includes various traffic variables, including harmonic average speed, free flow speed, historical speed, travel time, and annual average daily traffic (AADT) for each link of highways. INRIX is creating the NPMRDS dataset of average speeds and travel times for specific road links across the NPMRDS road network for every 5 min. INRIX uses its large data source to include millions of connected vehicles and trucks that are supplying location and movement data anonymously. They also implemented a path-processing algorithm to meet NPMRDS requirements using its existing source data.

For this project, the value of travel time from this dataset is chosen to be the target variable. This variable is defined as the ratio between the length of each link and the harmonic average speed of vehicles passing the link. Because the travel time of each link is dependent on the length of the link, the harmonic average speed is selected as the target. By predicting the average speed, the value of travel time could be calculated using the length of each link. The missing values of this dataset are imputed by temporal, spatial, and historical estimation.

Figure 10 shows a sample of the weekly average speed for the links of the eastbound and westbound Eisenhower Expressway. Daily traffic patterns, beginning Monday, September 11, 2017, to Sunday, September 17, 2017, are displayed. Weekdays have similar traffic patterns, although Saturday and Sunday are different. Figure 11 shows the daily speed for links of the eastbound and westbound Eisenhower Expressway. Figure 12 displays the distribution of speed on the eastbound and westbound Eisenhower Expressway for both links. Figure 13 shows a histogram of the average speed from this dataset.



(a)



(b)

Figure 10. Weekly average speed of Eisenhower Expressway links from probe-vehicle dataset (a) eastbound and (b) westbound (September 2017).

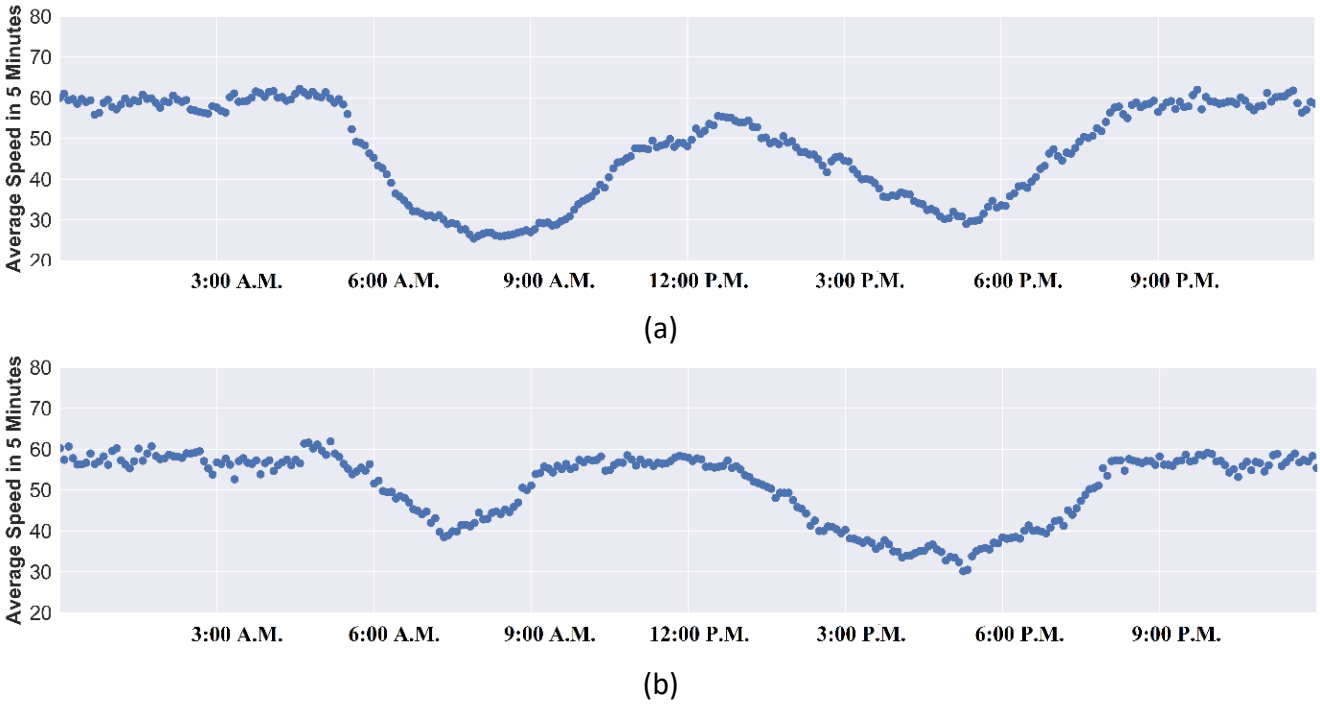


Figure 11. Daily average speed of Eisenhower Expressway links from probe-vehicle dataset (a) eastbound and (b) westbound (September 12, 2017).

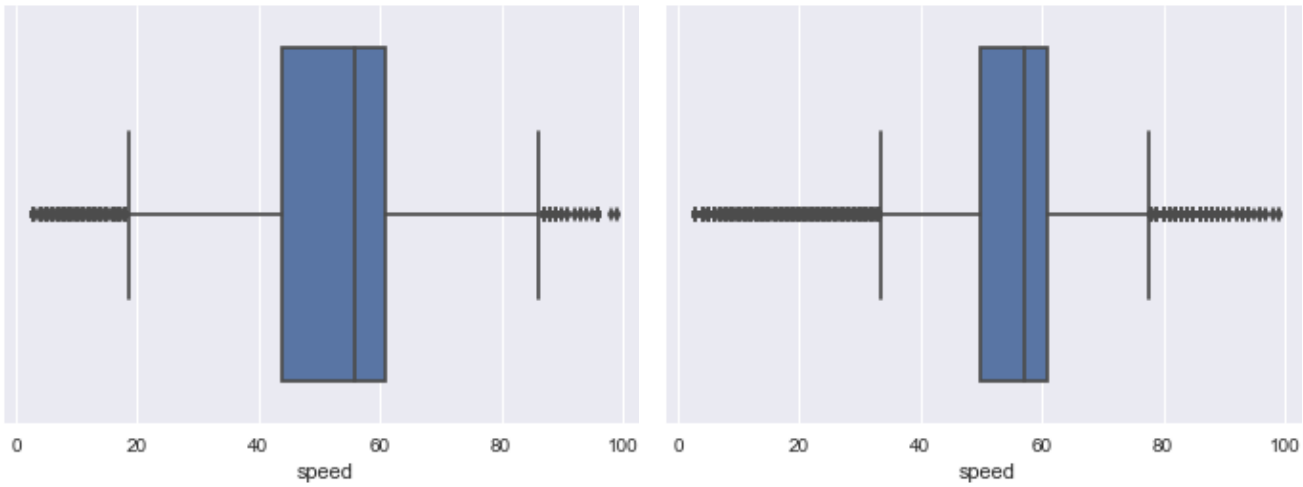


Figure 12. Distribution of speed for Eisenhower Expressway links from probe-vehicle dataset (a) eastbound and (b) westbound (September 2017).

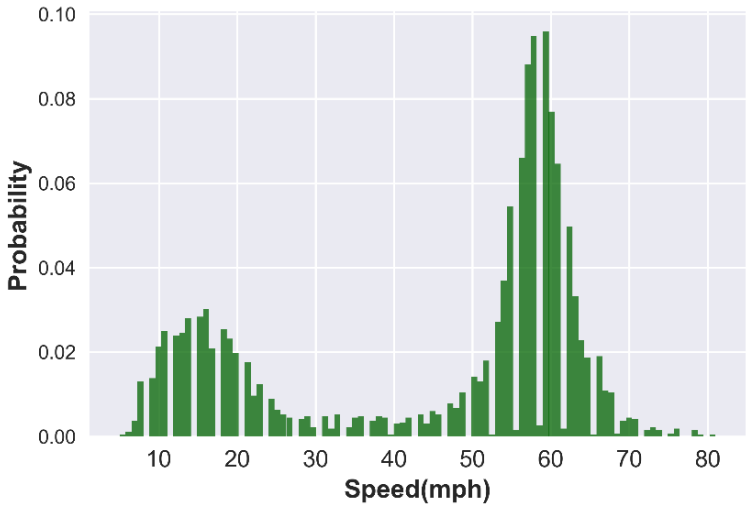


Figure 13. Average speed histogram from probe-vehicle dataset (September 2017).

Weather-Condition Data

Hourly weather-condition data, provided by the National Weather Service, is another source of data used in this study. This data is available for airport stations and based on the location of selected corridor data from Midway International Airport, which is the closest station utilized. In the dataset, weather condition is reported in 94 states. All states are grouped into four categories from fair to severe weather conditions. The first category contains weather conditions such as “fair” and “partly cloudy.” The second category is comprised of weather conditions such as “fair with light haze” and “light haze dust and windy.” Unfavorable weather conditions such as “light drizzle” and “light rain fog” constitute the third category. The fourth and most severe category includes weather conditions like “snow fog” and “thunderstorm heavy rain fog.” In addition to the weather condition, surface temperature is also available in the dataset used in this study. Like other data sources, there are missing values in this dataset, which are imputed using the information from previous and next time intervals.

Roadway Incident Data

Roadway incident data is also collected by the Gateway Traveler Information System, and details of roadway incidents occurring on highways are reported. Time of roadway incident, geographical location, and severity of roadway incident are among the variables in this dataset.

Based on an analysis of the roadway incident data, when a roadway incident occurs on a highway, that link, previous links, and the next links on the highway are affected, on average, for half an hour after the roadway incident. This effect is considered as binary variables to show the affected links and time intervals in the model. Because the effect of a roadway incident is different upstream and downstream, two binary variables for previous links and next links are defined. Another variable in the model also demonstrates roadway incident severity in three levels: minor, medium, and major. Whenever roadway incident severity is not determined at the roadway incident detection time, the medium level is selected, which is the most common. Figure 14 shows the percentage of severe roadway incidents in the dataset.

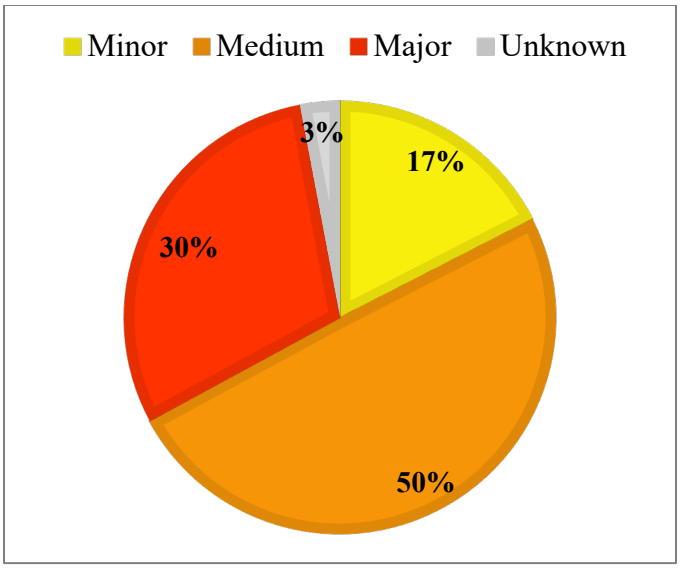


Figure 14. Percentage of severe roadway incidents in the dataset.

Roadwork Data

Roadwork data, which is collected by the Gateway Traveler Information System, contains details of construction or maintenance operations on highways. Variables such as start time, end time, geographical location, and severity of roadwork are reported in this dataset.

To consider the effect of this variable on the model, a binary variable is defined. When a roadwork project is executed on a link of a highway, that link, the previous link, and the next link are affected from start time to end time of the project. Also, if one of the lanes of the highway is blocked, then the variable "number of lanes" (which is from geometry data) considers its effect on the model. Figure 15 displays the severity of roadwork in the dataset.

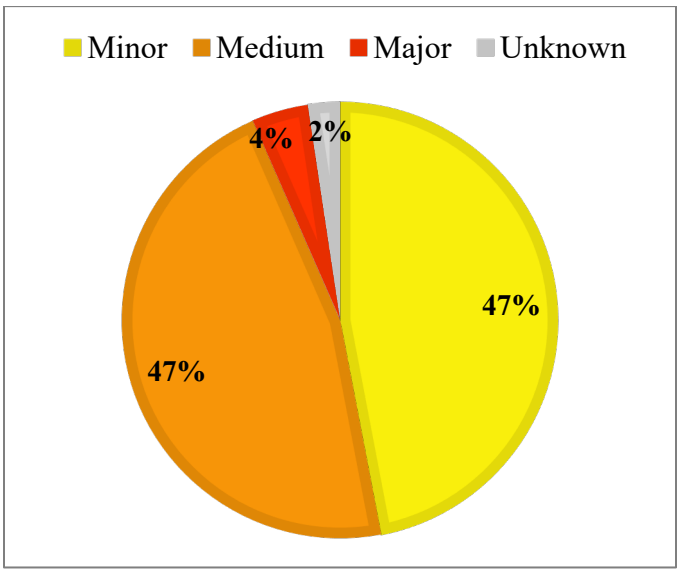


Figure 15. Severity of roadwork in the dataset.

Special-Event Data

The special-event dataset collected by the Gateway Traveler Information System provides information about events, including stadium events, parades, and road races. Several variables such as start time, end time, location, and type of event are reported in this dataset. Figure 16 shows the percentage of special-event types in the dataset.

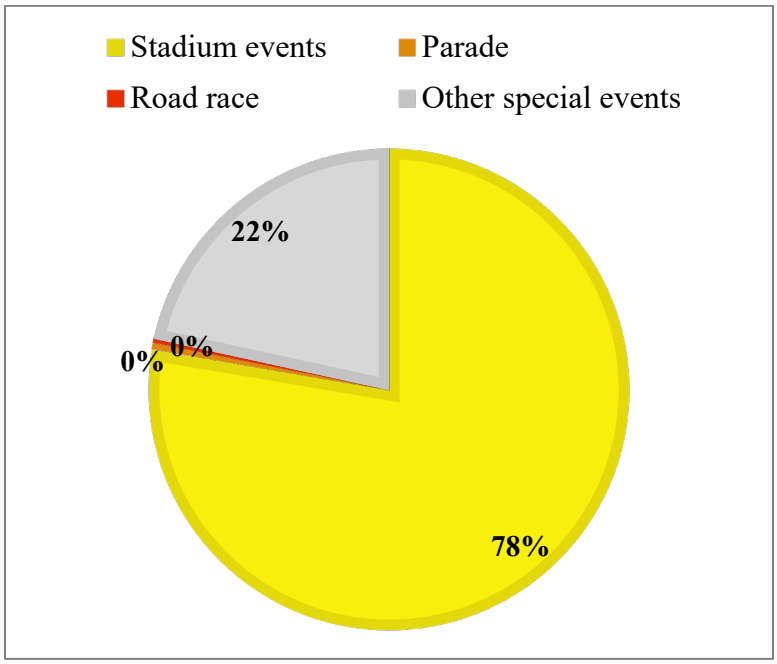


Figure 16. Percentage of special-event types in the dataset.

When a special event occurs close to highway links, the links are affected by that event for one hour before or after that event. The traffic one hour before or after an event depends on the direction of extra traffic and the relative location of the event and each link. If a highway link is located toward the path of an event location, extra traffic passes that link before the event, so one hour before the event is considered an effective time for that link. For links that are in the path from the event location toward other parts of the city, one hour after the event is the affected time.

Geometry Data

The geometry dataset includes variables from the studied highway corridor, including number of lanes, number of entrances, number of exits, and length of each highway link.

Sun-Glare Data

Sun glare is an environmental hazard that can cause crashes and impact traffic flow by increasing congestion. There are several definitions for glare according to available sources. Cline, Hofstetter, and Griffin (1997) define sun glare as, “relatively bright light, or the dazzling sensation of relatively bright light, which produces unpleasantness or discomfort, or which interferes with optimal vision.” Every year, many people are involved in crashes caused by sun glare. To this end, predicting accurate time, location, and impact of sun glare on highways is important. In this section, the methodology for

considering sun glare is provided. The exact time and locations influenced by sun glare are also specified, and the impact of sun glare on the travel time prediction model will be examined and discussed.

Many drivers have been "blinded" for several seconds while driving because of the low position of the rising and setting sun. However, it is not accurately known when the sun starts and stops impacting drivers' vision and traffic flow. The impact of sun glare also varies among drivers. Some drivers can drive more conveniently than others under the impact of sun glare, while other drivers cannot continue driving and brake or change their path. The condition of a vehicle's windshield can also intensify the impact of sun glare. A dirty or pitted windshield causes the sun to refract within the glass. This condition can get worse with a cracked or chipped window. Several researchers showed that sun glare can slow down traffic flow (Auffray et al. 2008; Churchill, Tripodis, and Lovell 2012). Auffray et al. (2008) showed that sun glare can affect traffic flow regardless of traffic condition (i.e., congested/uncongested). Churchill et al. (2012) also displayed the connection between sun glare and congestion in freeways, and their results showed that sun glare could slow down traffic and lead to traffic congestion. Jurado-Piña et al. (2010) proved that sun glare can be predicted using a geometrical model between the driving vehicle and the position of sun. They also found that the direction of streets is important, as southeast and southwest directions have a larger chance of having sun glare in one day. In the summer, streets toward the northeast and northwest may have sun glare in the morning and afternoon, respectively. In the winter, the position of the sun is more southward, so streets toward the northeast and northwest have no sunrise and sunset sun glare (Li et al. 2019).

The impact of sun glare on drivers' vision and, consequently, on traffic flow have been investigated. The Federal Highway Administration (FHWA) indicated that 15% of traffic congestion found on highways is attributable to the weather (Cambridge Systematics 2005). Driver visibility is a key factor in the formation of traffic congestion, as lower visibility increases the travel time delay by +12% (Goodwin 2002). Main sources of low visibility include sun glare as well as rain, snow, or fog (Shepard 1996).

To determine the impact of sun glare on drivers, researchers provided different algorithms with respect to the position of sun (Reda and Andreas 2004; Blanco-Muriel et al. 2001; Michalsky 1988), some of which are available online (Giesen 2018; NOAA n.d.). Li et al. (2019) employed Google Street View panorama images to estimate and predict the occurrence of sun glare, which provided important tools for drivers and traffic planners. There are several factors that should be considered in these algorithms such as the sun's angle in relation to cars, which is further influenced by solar declination, solar elevation, solar azimuth, driving direction, and slope. Solar declination is the angle between the rays of the sun and the plane of the earth's equator. Solar elevation is measured vertically from the point on the horizon to the object, and solar azimuth is measured clockwise from true north to the point on the horizon directly below the object. Figure 17 represents solar geometry.

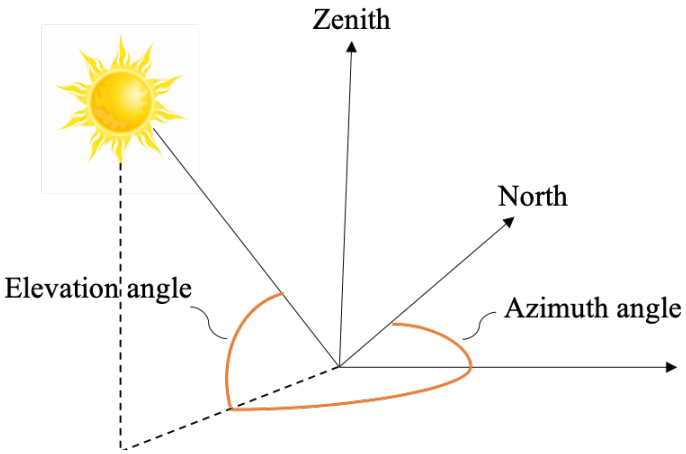


Figure 17. Solar azimuth and solar elevation.

Several factors can affect the determination of the angles' limits, including the height of drivers' eyes. The American Association of State Highway and Transportation Officials (AASHTO) provides the height of a driver's eye, which is 1080 mm above the road surface for passenger vehicles and 2330 mm for trucks (AASHTO 2001).

To find the locations and times that sun glare occurred on highways, the research team used SunEarthTools's website (n.d.) and took the thresholds of ± 5 and (0, 15) degrees for azimuth and elevation angles, respectively. Figure 18 shows an example of a location on the Eisenhower Expressway that meets the requirement of having a sun glare on September 1, 2017, at 18:14. The position of the sun on the mentioned date and time is within the predefined thresholds (relative solar elevation is 12.06 degrees and relative solar azimuth is 0.12 degrees), so that drivers who are moving west on the Eisenhower Expressway are exposed to direct sunlight.

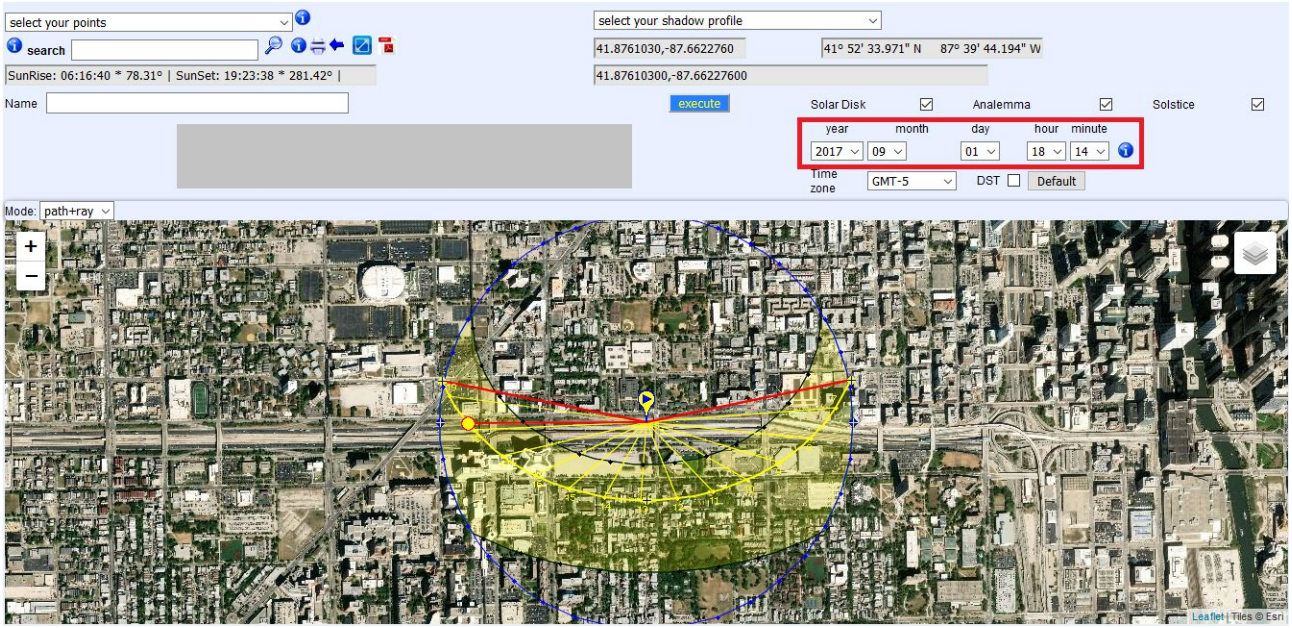
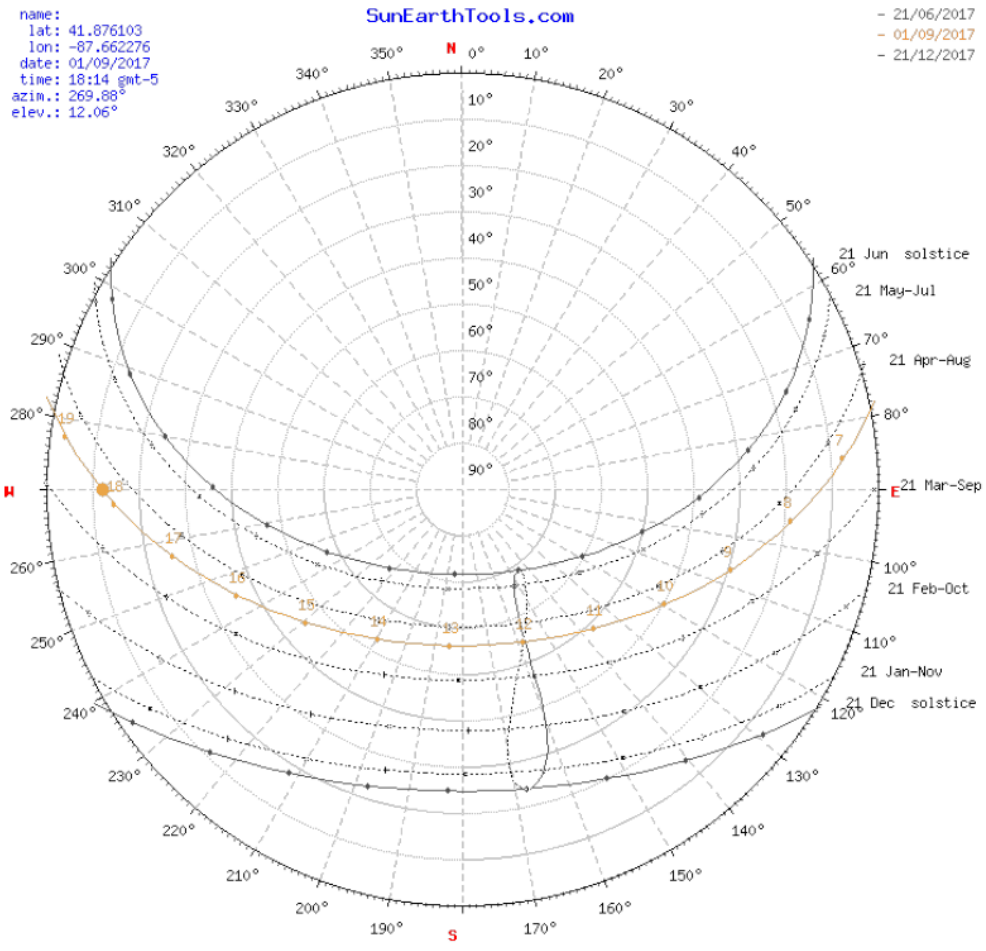


Figure 18. Sun position in a sun glare occasion on the westbound Eisenhower Expressway.

To determine the range of time in which drivers experience sun glare, refer to Figure 19, where the location of the sun (i.e., azimuth angle and elevation angle) is calculated with respect to different months and times. For example, the brown point in Figure 19 displays September 1, 2017, at 18:14. Based on the location of this point in the figure, we can determine that the elevation and azimuth angles are 12.06 degrees and 269.88 degrees, respectively. Based on the direction of the specific highway link, sun glare occurs when azimuth is between 265 to 275 degrees.

Finally, each link grade and direction are calculated, and then the azimuth and elevation angles are found using the "Pysolar" package in Python and NOAA's solar calculator (NOAA n.d.). The research team determined the exact links and times in the dataset where drivers experienced sun glare. Figure 20 presents a sample of data in which in each row of data an occurrence of sun glare is calculated as a binary variable. Therefore, Figure 20 shows that in the September 1, 2017, from 18:00 to 18:42 drivers who are driving in a specific link of the westbound Eisenhower Expressway (Figure 18) have been exposed to direct sunlight or sun glare.



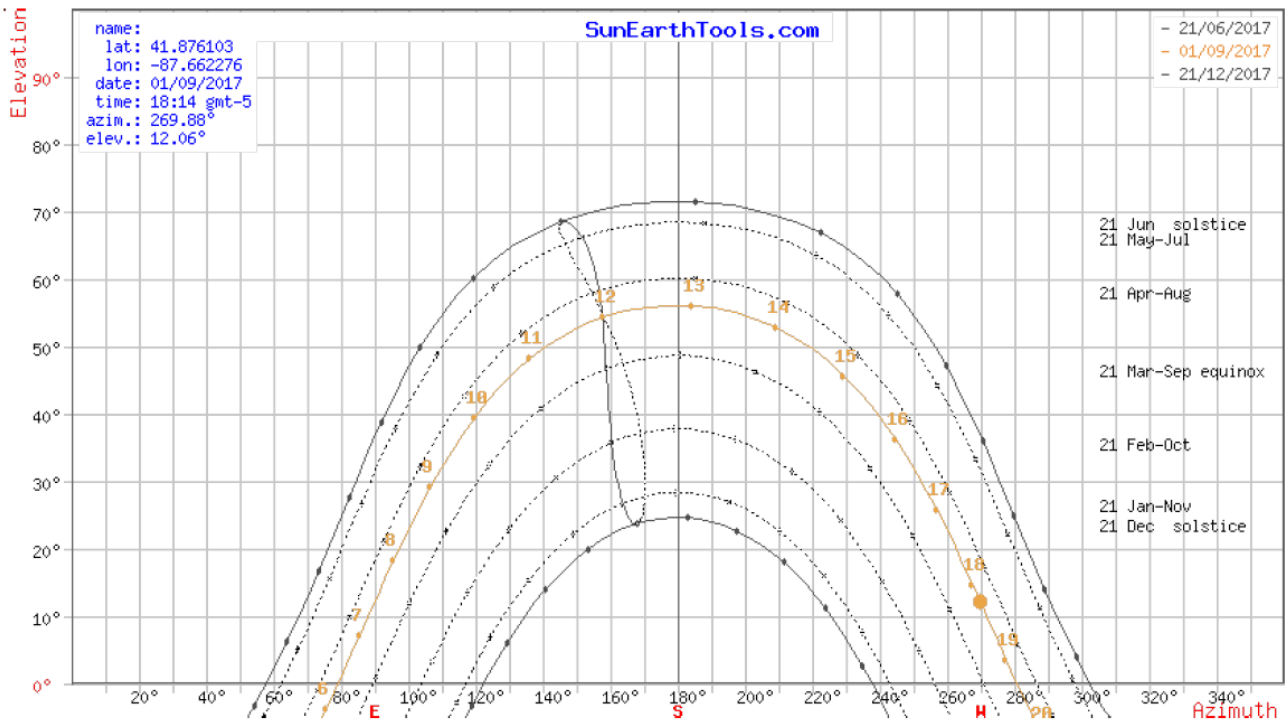


Figure 19. Sun position for a specific point on the Eisenhower Expressway.

Date	Time (past local midnight)	Julian Day	Julian Century	Geom Mean Long Sun (deg)	Geom Mean Anom Sun (deg)	Eccent Earth Orbit	Sun Eq of Ctr	Sun True Long (deg)	Sun True Anom (deg)	Sun Rad Vector (AUs)	Sun App Long (deg)	Mean Obliq Ecliptic (deg)	Obliq Corr (deg)	Sun Rt Ascen (deg)	Sun Declin (deg)	var y	Eq of Time (minutes)	HA Sunrise (deg)	Solar Noon (LST)	Sunrise Time (LST)	Sunset Time (LST)	Sunlight Duration (minutes)	True Solar Time (min)	Hour Angle (deg)	Solar Zenith Angle (deg)	Solar Elevation Angle (deg)	Approx Atmospheric Refraction (deg)	Solar Elevation corrected for atm refraction (deg)	Solar Azimuth Angle (deg cw from N)	sunglare-evening
9/1/2017	17:54:00	2457998.45	0.17668595	161.297	6718.06	0.0167	-1.60602	159.691	6716.45	1.00904	159.682	23.437	23.4349	161.236	7.9377	0.04302	0.195824	98.3206	12:50:25	6:17:08	19:23:42	786.565	1023.59	75.897	74.2241	15.7759	0.05628	15.8322	266.529	0
9/1/2017	18:00:00	2457998.46	0.17668606	161.301	6718.06	0.0167	-1.6061	159.695	6716.45	1.00903	159.686	23.437	23.4349	161.24	7.93619	0.04302	0.197158	98.3192	12:50:25	6:17:08	19:23:41	786.5538	1029.59	77.3973	75.3407	14.6593	0.06063	14.7199	267.545	1
9/1/2017	18:06:00	2457998.46	0.17668617	161.305	6718.06	0.0167	-1.60617	159.699	6716.46	1.00903	159.69	23.437	23.4349	161.244	7.93467	0.04302	0.198492	98.3178	12:50:25	6:17:08	19:23:41	786.5425	1035.59	78.8976	76.4582	13.5418	0.06565	13.6075	268.555	1
9/1/2017	18:12:00	2457998.47	0.17668629	161.309	6718.07	0.0167	-1.60625	159.703	6716.46	1.00903	159.694	23.437	23.4349	161.247	7.93315	0.04302	0.199825	98.3164	12:50:24	6:17:09	19:23:40	786.5312	1041.59	80.398	77.5762	12.4238	0.07149	12.4953	269.559	1
9/1/2017	18:18:00	2457998.47	0.17668640	161.313	6718.07	0.0167	-1.60632	159.707	6716.47	1.00903	159.698	23.437	23.4349	161.251	7.93163	0.04302	0.201159	98.315	12:50:24	6:17:09	19:23:40	786.52	1047.59	81.8983	78.6943	11.3057	0.07837	11.3841	270.559	1
9/1/2017	18:24:00	2457998.48	0.17668652	161.317	6718.08	0.0167	-1.60639	159.711	6716.47	1.00903	159.702	23.437	23.4349	161.255	7.93011	0.04302	0.202493	98.3136	12:50:24	6:17:09	19:23:40	786.5087	1053.59	83.3986	79.8122	10.1878	0.08658	10.2743	271.556	1
9/1/2017	18:30:00	2457998.48	0.17668663	161.321	6718.08	0.0167	-1.60647	159.715	6716.47	1.00903	159.706	23.437	23.4349	161.259	7.92859	0.04302	0.203827	98.3122	12:50:24	6:17:09	19:23:39	786.4975	1059.6	84.899	80.9296	9.07035	0.09654	9.1669	272.549	1
9/1/2017	18:36:00	2457998.48	0.17668674	161.325	6718.08	0.0167	-1.60654	159.719	6716.48	1.00903	159.71	23.437	23.4349	161.262	7.92707	0.04302	0.205161	98.3108	12:50:24	6:17:10	19:23:39	786.4862	1065.6	86.3993	82.0462	7.9538	0.10883	8.06263	273.541	1
9/1/2017	18:42:00	2457998.49	0.17668686	161.329	6718.09	0.0167	-1.60661	159.723	6716.48	1.00903	159.714	23.437	23.4349	161.266	7.92555	0.04302	0.206496	98.3094	12:50:24	6:17:10	19:23:38	786.475	1071.6	87.8996	83.1616	6.83844	0.12427	6.96271	274.531	1
9/1/2017	18:48:00	2457998.49	0.17668697	161.333	6718.09	0.0167	-1.60669	159.727	6716.49	1.00903	159.718	23.437	23.4349	161.27	7.92403	0.04302	0.20783	98.308	12:50:24	6:17:10	19:23:38	786.4637	1077.6	89.4	84.2754	5.7246	0.14405	5.86865	275.521	0
9/1/2017	18:54:00	2457998.50	0.17668709	161.338	6718.1	0.0167	-1.60676	159.731	6716.49	1.00903	159.722	23.437	23.4349	161.274	7.92251	0.04302	0.209164	98.3066	12:50:24	6:17:10	19:23:37	786.4524	1083.6	90.9003	85.3874	4.61262	0.16982	4.78244	276.511	0
9/1/2017	19:00:00	2457998.50	0.17668720	161.342	6718.1	0.0167	-1.60684	159.735	6716.49	1.00902	159.726	23.437	23.4349	161.278	7.921	0.04302	0.210499	98.3051	12:50:24	6:17:11	19:23:37	786.4412	1089.6	92.4006	86.4972	3.50282	0.20719	3.71001	277.501	0
9/1/2017	19:06:00	2457998.50	0.17668731	161.346	6718.1	0.0167	-1.60691	159.739	6716.5	1.00902	159.73	23.437	23.4349	161.281	7.91948	0.04302	0.211833	98.3037	12:50:24	6:17:11	19:23:37	786.4299	1095.6	93.901	87.6045	2.39554	0.25961	2.65515	278.494	0
9/1/2017	19:12:00	2457998.51	0.17668743	161.35	6718.11	0.0167	-1.60698	159.743	6716.5	1.00902	159.734	23.437	23.4349	161.285	7.91796	0.04302	0.213168	98.3023	12:50:24	6:17:11	19:23:36	786.4187	1101.61	95.4013	88.7089	1.29113	0.33688	1.628	279.488	0

Figure 20. Sun-glare calculation using NOAA's solar calculator.

FINAL DATASET

After applying the data-preparation process to each dataset, one of the most challenging parts is to combine the datasets properly. Each dataset has its own timestamp and location type. Because the goal is to analyze link-based data, locations from the probe-vehicle dataset (highway links) are used as base locations. Then, flow and occupancy data from the loop-detector dataset are added to the average speed values of the probe-vehicle dataset. For this purpose, the location of each loop detector on the highway is determined and compared to the location of highway links based on the probe-vehicle dataset. So, each loop detector located on a highway link is specified and the flow and occupancy of that loop detector are assigned to that link on each timestamp. In addition, flow and occupancy values of the loop-detector dataset are aggregated in 5 min intervals, which is like time intervals for average speed values reported in the probe-vehicle dataset.

The next step is to add the weather data to the dataset. The weather-condition dataset contains weather-related variables for the locations of airports. So, for each link of the highway, the distances of the link to airports are calculated, and the weather-condition variable for the closest airport is assigned to that link for each timestamp. Based on the location of the studied corridor, information from Midway International Airport is selected for all links in this study.

Roadway incident cases are also added to the dataset based on their location and timestamp. First, the closest highway link is determined based on the geographical location reported for each roadway incident. Then, the roadway incident effect is considered for all timestamps based on the observation of different roadway incident reports and their effects on highway traffic. The roadway incident effect is the time of detection to half an hour after that time on the specified link as well as the five previous and five next links on the highway.

The effects of roadwork in the model are considered based on their reported geographical location. All links located in the roadwork zone are determined, and the effect of roadwork is considered for all links on the timestamps from the roadwork's start to end time. Information on network geometry that is assigned to each link of highways is generally constant over time. The only variable that is not constant over time is number of lanes, which might change during roadwork times. So, the geometry information for each link is added to the dataset.

The effect of special events on highway traffic depends on the location of an event. To help improve the model's accuracy, there is an important special-event variable to consider along with traffic-related variables when driving westbound on the Eisenhower Expressway. This variable is the United Center. Events held at the United Center have significant impact on the traffic of the Eisenhower Expressway before they start and after they finish. Based on the location of the United Center and available routes, some highway links are located in routes toward the arena, which are affected before events. Figure 21 shows the location of these links with a blue arrow. However, some links located in routes from the arena are affected after events, which are shown by the orange arrow in Figure 21. So, the special-event data for all these links at their effective time is added to the main dataset. Note that the effect of all events on highway traffic are considered indirectly by traffic-related variables in the model.

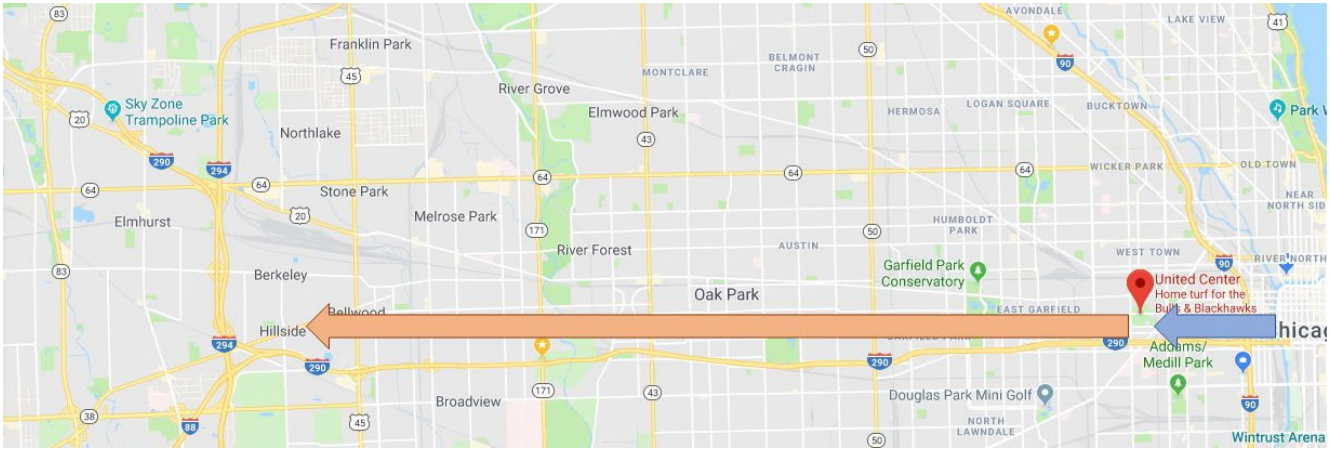


Figure 21. Location of links on the Eisenhower Expressway affected by events at the United Center.

We specified the links that are impacted by special events through an analysis of special events and their impacts on traffic condition. Table 2 displays parts of highways in which different impacts of special events should be considered.

Table 2. Location of Special Events and Affected Parts of Highways

Location	Highway	Before Events		After Events	
		Direction	Link Numbers*	Direction	Link Numbers
Soldier Field, Northerly Island, Grand Park, and Navy Pier	Dan Ryan	E	65–78	W	20–31
	I-55	N	570–579	S	8–17
	I-290	E	60–69	-	-
	Dan Ryan	W	15–20	E	80–85
Wrigley Field	Kennedy	E	26–33	W	66–72
	Kennedy	W	55–64	E	36–44
United Center	I-290	E	50–58	W	10–17
	Kennedy	E	53–57	W	41–46
	I-290	W	1–12	E	57–69
Guaranteed Rate Field	Dan Ryan	E	76–85	W	12–24
	I-55	N	569–572	S	13–16
	I-55	S	8–12	N	573–579
	Dan Ryan	W	6–14	E	84–92

* Link numbers are based on the definition of the probe-vehicle dataset.

In the final dataset, each row of data contains information of a highway link at a 5-min time interval. Some variables such as number of exits or length of link are constant over time, some variables are reported every hour such as temperature, and traffic-related variables must be updated every time interval (5 min). Note that selecting 5 min for aggregation is reasonable because it is applied in similar studies (Yildirimoglu and Geroliminis 2013; Chen and Chien 2001; Chen, Rakha, and McGhee 2013; Hussein, Brown, and Herrmann 2019). A 5-min time period does not reflect unnecessary traffic fluctuations and is able to consider significant changes in traffic state. Table 3 shows variables of the final combined dataset, with more than 3 million records, used to develop travel time prediction models.

Table 3. Description of Explanatory Variables for Travel Time Prediction Models

Variable	Description
Timei i = 1, 2, …, 8	1: if time ID equals i*, and 0: otherwise
Friday	1: if the day of week is Friday, and 0: otherwise
Saturday	1: if the day of week is Saturday, and 0: otherwise
Sunday	1: if the day of week is Sunday, and 0: otherwise
WeatherID	Weather condition: 1: Fair or a few clouds 2: Cloudy with haze or light fog 3: light rain or drizzling 4: Heavy rain, snow, thunderstorm or fog
Temp_F	Temperature (Fahrenheit)
Occupancy	Aggregated occupancy at location of loop detector which is closest to midpoint of the link at the 5-min time interval

Variable	Description
Occupancy_p	Aggregated occupancy at location of previous loop detector on the road at the 5-min time interval
Occupancy_n	Aggregated occupancy at location of next loop detector on the road at the 5-min time interval
Count	Aggregated count of vehicles on loop detector which is closest to midpoint of the link at the 5-min time interval
Count_p	Aggregated count of vehicles on previous loop detector on the road at the 5-min time interval
Count_n	Aggregated count of vehicles on next loop detector on the road at the 5-min time interval
AADT	Annual average daily traffic of the link
EntRamp	Number of entrance ramps on the link
ExtRamp	Number of exit ramps on the link
Miles	Length of the link in miles
Lanes	Number of lanes on the link
IsAcc_p	1: if a roadway incident occurred on the link or previous links, at the 5-min time interval, and next half-an-hour time intervals after accident occurrence, and 0: otherwise
IsAcc_n	1: if a roadway incident occurred on the link or next links at the 5-min time interval, and next half-an-hour time intervals after accident occurrence, and 0: otherwise
AccSeverityID	Roadway incident severity: 1: Minor severity 2: Medium severity 3: Major severity
IsRW	1: if a roadwork project is executing on the link, previous link, or next link at the 5-min time interval, and 0: otherwise
IsSE	1: if a special event is occurring close to the link at the 5-min time interval, and either next one-hour, or previous one-hour time intervals**, and 0: otherwise
IsSunGlare	1: if sun glare is occurred on the link at the 5-min time interval, and 0: otherwise
Speed (target)	Average speed of vehicles on a link at the 5-min time interval

* Each 3-hr time interval of a day is considered a binary variable. For instance, Time1 represents an interval from midnight to 3 a.m., and Time8 represents an interval of 9 p.m. to 12 a.m.

** Considering the next or previous hour depends on the direction of extra traffic and the relative locations of the event and link. For example, if a link is toward the path of an event, extra traffic passes this link before the event. In this case, we consider the previous hour as the effective time.

CHAPTER 4: MODELS AND TECHNIQUES

Machine learning methods are compatible with the type of data used for this study to develop the travel time prediction model. K-nearest neighbors (KNN), backpropagation neural network (BPNN), and random forest (RF) models are selected based on their specific properties. In addition, K-means and fuzzy C-means clustering methods are selected to categorize data points in order to investigate whether clustering before applying prediction models could improve the accuracy of results.

The goal was to cluster data points in order to locate all homogeneous data points in each cluster. Then, KNN, BPNN, and RF models are developed for each cluster. The appropriate cluster is first determined for each new cluster. Then, based on the model corresponding to that specific cluster, the speed value is predicted using the best model. If the fuzzy clustering method is used, then the speed values for all models of different clusters are predicted. The result would be the weighted average of the predicted speed values, in which the weights are the membership grades of the fuzzy clusters. Accuracy is expected to increase by developing a specific model for each cluster.

TRAVEL TIME PREDICTION METHODS

K-nearest Neighbors

The K-nearest neighbors model only needs a large volume of data points without developing a mathematical model that could predict the target variable. There is also no need to define the parameters of the model in advance. The authenticity of the data is maintained because no smoothing procedure is done to the data. This method is an appropriate choice for the nonparametric problem of travel time prediction (Yu et al. 2016).

K-nearest neighbors is a supervised machine learning technique that is widely used for regression and classification. The KNN method classifies a dataset by comparing similar records of a trained dataset (Han et al. 2011). The KNN regression algorithm is used in this study, because the target variable is speed, which is a continuous value rather than a predefined class. In general, the training dataset has n attributes, and each of its records could be presented in a n-dimensional variable space by the value of the n attributes. After that, to predict the target value for new records from the test dataset, each new record with its attribute values finds its location in the variable space. The algorithm looks for the k-nearest neighbors that are closest to the new record among the trained data points. In this study, the speed value for a new record would be predicted by taking the average of the speed values of K-nearest neighbors to that record. In other words, the method searches for nearest neighbors among all historical data points and uses them for the prediction.

The Euclidean metric is used to calculate the distance between two points in the n-dimensional variable space. Between two points of $X_1 = (x_{11}, x_{12}, \ldots, x_{1n})$ and $X_2 = (x_{21}, x_{22}, \ldots, x_{2n})$, the Euclidean distance is defined in Equation 4-1:

$$\text{Euclidean distance } (X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \qquad (4\text{-}1)$$

Before calculating the distance between points, the attributes' values are normalized to prevent attributes with larger values, such as occupancy, from outweighing attributes with smaller values, such as binary attributes (Han, Pei, and Kamber 2011). Figure 22 demonstrates the schematic diagram of the KNN model. To predict the speed value of the red point in a two-dimensional variable space with k equal to 5, all points within the green circle are defined as neighbors of the red point. So, predicted speed value of the red point is calculated by getting the average of the neighbors' speed values.

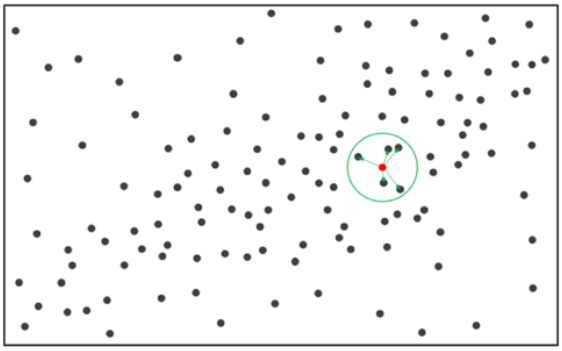


Figure 22. Schematic diagram of the KNN model.

Backpropagation Neural Network

Backpropagation neural network (BPNN) is a popular supervised machine learning method that is inspired by the biological nervous system. Each neural network has three main layers: one input layer, one or more hidden layer(s), and one output layer. Each layer consists of one or more neuron(s). The neural network is trained to adjust weights to the connection between neurons of different layers in order to perform a specific function. A neural network is trained by training data points to find the weights of network connections in order to connect a specific input to a desired output. The value of each neuron *p* is multiplied by the connection weight *w* and added to a bias *b*, and the result is *wp+b*. This value, which is the input for the transfer function *f*, results in the *f* output *f(wp+b)*. The values of *w* and *b* will be adjusted to find the desired output values. The backpropagation method, which recalculates and minimizes error values automatically, is utilized in this project.

Random Forest

Random forest is an ensemble supervised machine learning method that can be used as a classifier or regressor for categorical or numerical datasets (Han, Pei, and Kamber 2011). Several random decision trees (DT) are used in this method, so each DT votes to predict the value of the target variable. To combine DTs, bootstrap aggregation (or bagging) is used, which is a model averaging approach for combining machine learning methods to increase accuracy. To develop a random forest model, multiple random samples from the training dataset are selected with replacement in several iterations. A DT is trained for each random sample. Then, for a new record from the test dataset, each trained DTs returns the value of the target variable. The final result is calculated by taking the average of all predicted values for the target from DTs. Random forest is robust to noisy data and overfitting. It is expected to have higher accuracy than an individual DT, because it decreases variance (Han, Pei, and Kamber 2011). Random forest typically works accurately and quickly when a large

dataset is available. It can also manage as many variables as the model inputs. These characteristics make the random forest model an appropriate choice for predicting travel time (Fan et al. 2018).

Clustering Methods

K-means Clustering

The K-means clustering method is a popular technique in data mining. It partitions data points into k clusters so that each data point is grouped into the cluster with the nearest mean. The objective function aims to maximize similarity of data points within each cluster, while it minimizes similarity of data points from different clusters. Generally, the quality of each cluster is measured by calculating the within-cluster variation, which is the sum of the squared error between all data points in cluster C_i and the centroid c_i. Equation 4-2 demonstrates the sum of the squared error between all data points in the dataset, where p is a multidimensional data point. Because it is challenging to optimize this objective function, the K-means approach could be used.

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist\ (p, c_i)^2 \tag{4-2}$$

The K-means algorithm starts with k random points as initial clusters' centers. Then, each data point is assigned to a similar cluster so that the distance from the cluster center to that data point is minimized. There are different distance metrics such as Euclidean distance (Equation 4-1), which is used to calculate the distance between two points. Next, the new mean value for each cluster is calculated by obtaining the average of all members. Continue assigning data points to clusters and updating the cluster means until the mean values do not change. Therefore, this algorithm clusters data points by iteratively improving the value of within-cluster variation (Han, Pei, and Kamber 2011).

Fuzzy C-means Clustering

Fuzzy clustering, or soft clustering, is a clustering method in which each data point is grouped into more than one cluster with different membership grades. The membership grade shows the degree to which each data point belongs to each cluster. So, data points that are on the edge of each cluster might be in that cluster to a lesser degree than points that are located in the center. For a dataset with l data points, u_{ik} represents the membership degree of point p_k belonging to cluster C_i. This method aims to minimize the objective function, which is defined in Equation 4-3:

$$f = \sum_{k=1}^{l} \sum_{i=1}^{c} (u_{ik})^m dist(p_k\ , c_i)^2 \tag{4-3}$$

Subject to: $u_{ik} \in [0,1]\ and\ \sum_{i=1}^{c} u_{ik} = 1$ for all k, where c is number of clusters, c_i is center of cluster C_i, and $m > 1$ is a constant.

The problem could be solved by iteratively improving Equations 4-4 and 4-5.

$$u_{ik} = \left[\sum_{j=1}^{c} \left(\frac{dist(p_k, c_i)^2}{dist(p_k, c_j)^2} \right)^{\frac{1}{m-1}} \right]^{-1} \tag{4-4}$$

$$c_i = \frac{\sum_{k=1}^{l} (u_{ik})^m p_k}{\sum_{k=1}^{l} (u_{ik})^m} \tag{4-5}$$

CHAPTER 5: RESULTS

The methods explained in Chapter 4 are applied to the westbound Eisenhower Expressway, and the results are presented in this chapter. A similar procedure is applied for other highways in each direction, and the results are presented in this chapter as well.

CLUSTERING MODELS

After applying the data-cleaning process on different data sources and combining them to create the final dataset, 80% of the data points are randomly selected for model training. The remaining 20% are used as test data to evaluate the models. Then, K-means and fuzzy C-means clustering methods are applied based on the occupancy and flow variables. The best number of clusters is three for both methods. Figure 23 indicates the data points in three clusters based on the K-means method.

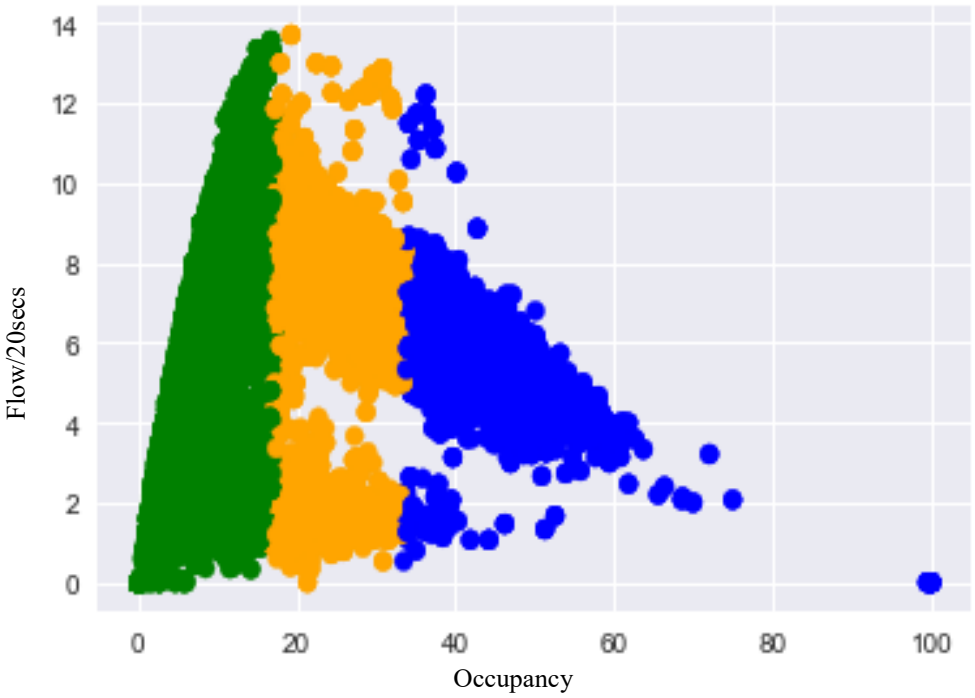


Figure 23. Clustered data points using the K-means clustering method.

Figure 24 shows data points that are clustered using the fuzzy C-means method and have membership grades more than 80% and 50% in the three defined clusters. In Figure 24-A, many data points on the edge of the clusters have membership grades less than 80%, so they are not shown in the figure.

The clustered data points using the K-means and fuzzy C-means models are provided for use after selecting the travel time prediction model.

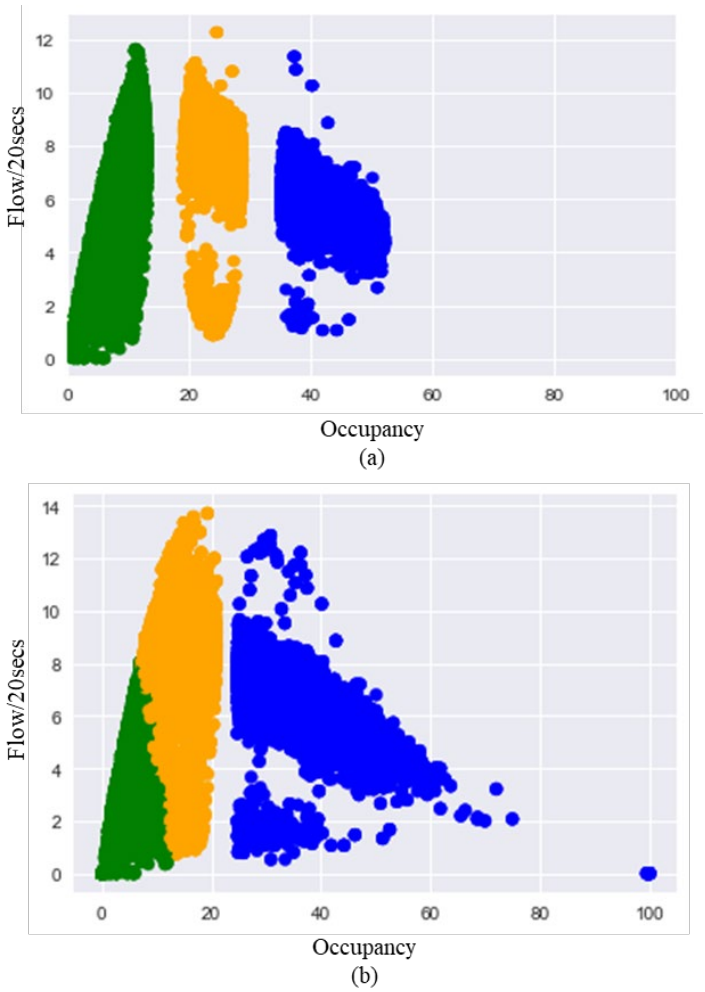


Figure 24. Clustered data points using the fuzzy C-means clustering method: membership grades more than (a) 80% and (b) 50%.

TRAVEL TIME PREDICTION MODELS

We used harmonic speed as the target variable. However, because the objective is to predict travel time, we can divide the highway's link length by predicted speed to determine travel time.

Before starting the model-training process, input variables are normalized. The number of neighbors for the KNN model is assigned experimentally as 4. For the random forest model, 20 DTs in a forest with a max depth of 20 are set. The neural network model has a lower accuracy than KNN and random forest models, so the results focused on the two better models—KNN and random forest. Finally, the KNN and random forest models are trained for the entire dataset. Then, they are applied to clustering methods to predict and compare short-term average speed as a representative of travel time for 5-min prediction horizons one hour ahead. That is, a model is trained for each technique to predict travel time at 0 min, 5 min, 10 min, and 60 min ahead.

Figure 25 displays and compares the prediction accuracy (i.e., R-squared score) of KNN and random forest models. The random forest model outperforms the KNN model in all 13 prediction time

intervals. However, the difference between their prediction accuracy is more negligible for short-term prediction horizons. Both models perform well at 0 min ahead (i.e., at the prediction moment). However, random forest slightly outperforms KNN with prediction accuracies of 95.3% and 92.9%, respectively. Prediction accuracy decreases by increasing the prediction horizon. Prediction accuracies at 5 min and 10 min ahead for the random forest model are 93.9%, and 91.6%, respectively. Prediction accuracies at 5 min and 10 min for the KNN model are 91.1%, and 88.1%, respectively. Based on Figure 25, prediction accuracy of both models is decreasing at a faster rate 30 min ahead of the horizon than 30 min to 60 min ahead. Despite the plateauing effect for prediction horizons above 30 min for both models, the performance of the random forest model is more robust. It achieves a prediction accuracy of 84% at 60 min ahead, while the KNN model achieves only 76.8%.

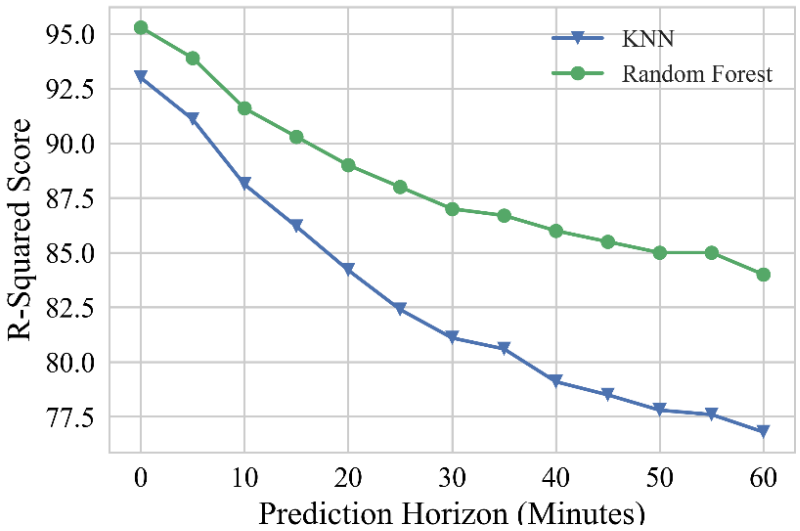


Figure 25. Accuracy of KNN and random forest models for different prediction horizons.

Based on Figure 25, both models, especially random forest, are well-suited to predict travel time in a short-term horizon (i.e., 15 min ahead). To better observe short-term prediction performance, true speed values versus predicted speed values of KNN and random forest models are plotted in Figure 26. This figure confirms that the shorter the prediction horizon, the more accurate the prediction for both models. Note that random forest outperforms KNN regarding training time.

Feature importance analysis is applied to the random forest model, and Table 4 shows the results. As expected, traffic variables most affect travel time. After traffic variables, temperature is found to be effective in the model, which shows the importance of weather conditions on traffic-related predictions. Unlike our expectation, roadway incidents, special events, and roadwork incidents have a relatively lower impact on the model. However, the low importance of these variables might stem from a small number of these incidents in the data. Note that because the effect of these incidents is highly correlated with traffic variables, which are important inputs of the model, their effects are considered indirectly through traffic variables in the model. Also, adding the sun glare variable to the models could not improve the accuracy significantly. Although sun glare occurrence could reduce speed, the main impact of sun glare is captured indirectly by other traffic-related variables used in the model.

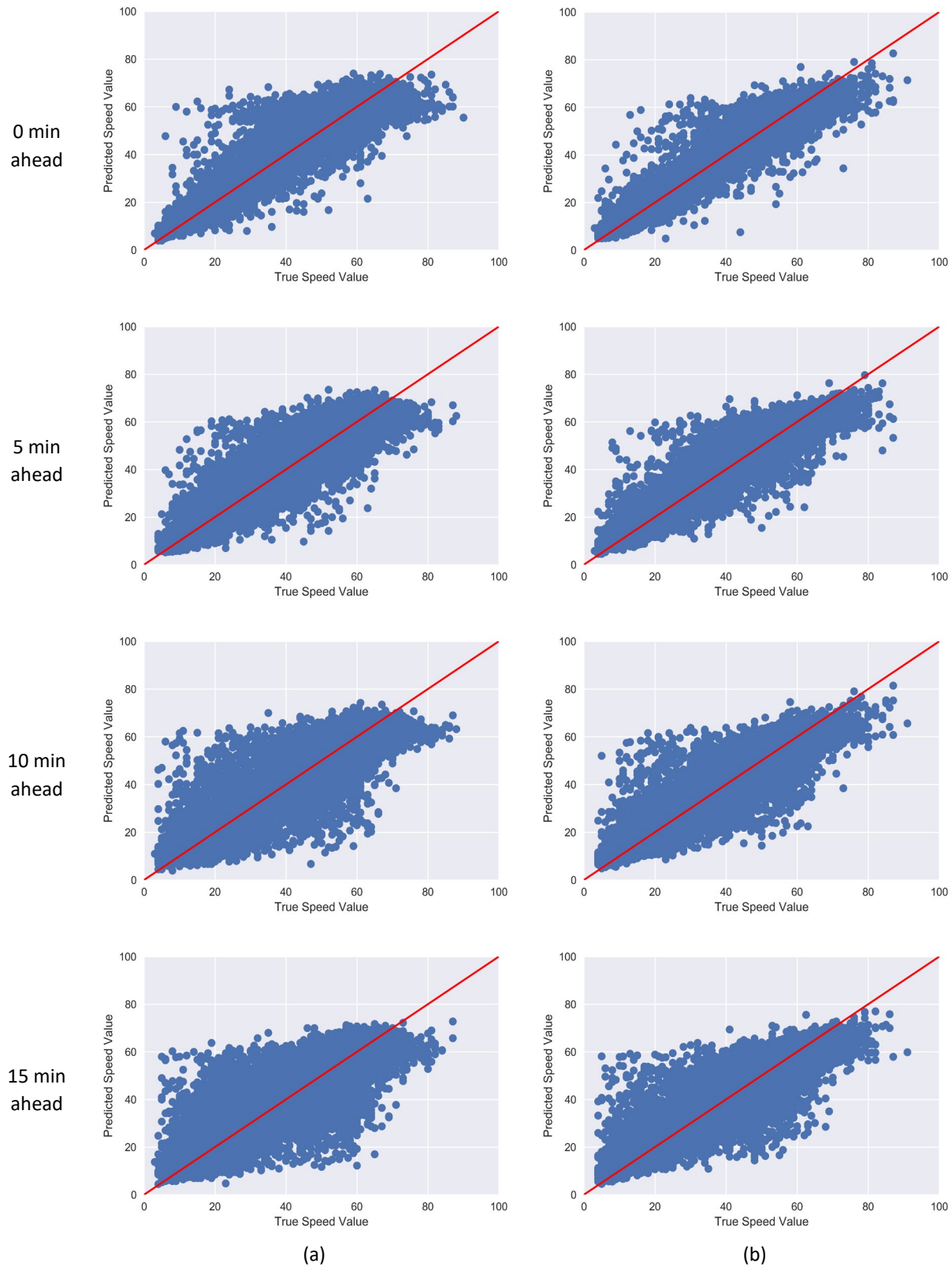


Figure 26. True speed values versus predicted speed values: (a) KNN and (b) random forest.

Table 4. Feature Importance Analysis Results

Variable	Rank
Occupancy	1
Count_p	2
Occupancy_p	3
Count	4
Occupancy_n	5
Count_n	6
AADT	7
Temp_F	8
Miles	9
Time3 (Morning Peak-Hours)	10
Time7 (Evening Peak-Hours)	10
Sunday	11
Saturday	12
Friday	13
ExtRamp	14
Timei (i = 1, 2, 4, 5, 6)	15
EntRamp	16
AccSeverityID	17
WeatherID	18
Lanes	19
Evets (IsAcc, IsRW, IsSE), Sun glare	20

The feature importance analysis could only relatively rank the importance of variables in the model. Therefore, a sensitivity analysis is conducted as well to assess the potential impact of important variables on the target variable. The top three important variables in the list are selected: occupancy (Occupancy, Occupancy_p, and Occupancy_n), count (Count, Count_p, and Count_n), and AADT. Then, the final random forest model is run several times with changes to each variable by ±10%, ±20%, ±30%, ±40%, and ±50%, while other variables remain the same. Figure 27 demonstrates the sensitivity of average speed for all links and time intervals to important variables. Based on the figure, increasing the occupancy reduces the average speed, and a linear relationship is expected. A nonlinear relationship is observed for the count variable. As mentioned, the average speed value is for all links and times under different conditions. On average, the speed increases by increasing the count from -25% to +50% and by decreasing the count from -25% to -50%. The effect of AADT on average speed is also nonlinear. By increasing AADT, the average speed also increases, but there is no significant change in average speed by increasing AADT from +10% to +50% and from -50% to -10%.

For data points in each cluster, KNN and random forest models are developed to investigate the performance of models using clustered data points. Contrary to expectation, the accuracy of the new models decreased for all conditions. Higher accuracy is achieved for the models that are trained using the entire dataset before applying clustering methods. So, the random forest model without clustering is selected for travel time prediction.

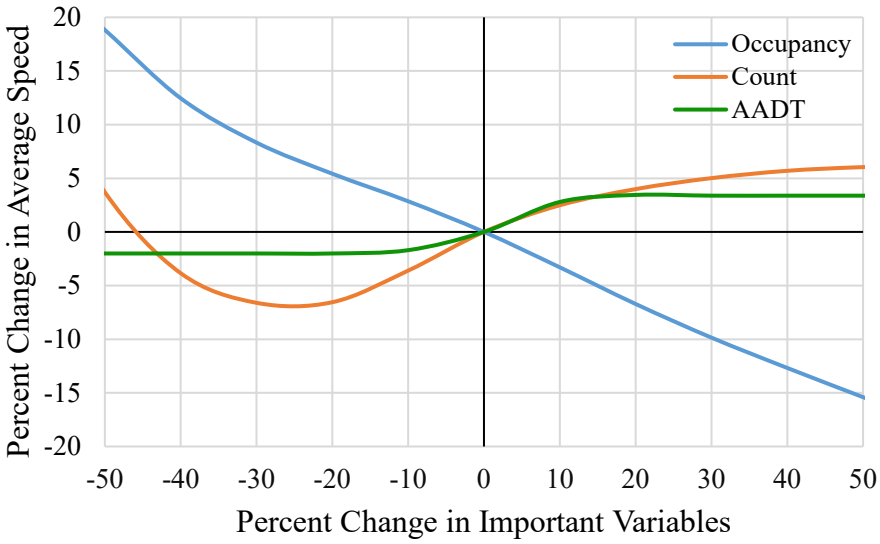


Figure 27. Sensitivity analysis of important variables in the travel time prediction model.

DYNAMIC TRAVEL TIME PREDICTION APPROACH

Based on the achieved results, the random forest model is selected as the best method in this project to estimate and predict travel time. By utilizing the input variables for each link of a highway, we can estimate the travel time for the moment and predict travel time at 5-min intervals. A dynamic travel time prediction approach is employed to achieve the highest accuracy.

In this approach, all links that form the route are defined to predict travel time for a specific route. Figure 28 shows a sample westbound route from the Eisenhower Expressway. Point A is the origin, and point B is the destination. The goal is to predict the specific travel time for this route, which is formed by 25 links.

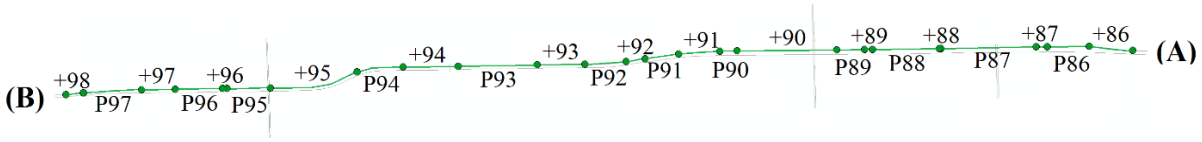


Figure 28. A sample route from the Eisenhower Expressway.

For each link, travel time is predicted at 5-min intervals. Average travel time is around 15 min, so we predict travel time at four 5-min intervals, or 20 min ahead. We assume that the current time is 7:00 a.m. We start at the first link and check travel time for the next 5-min time interval (from 7:00 a.m. to 7:05 a.m.). If travel time is less than 5 min, then we use the same time interval for the next link. We continue for the next links until aggregated travel time from the origin is more than 5 min. Then, for the next links, we use the travel time of the next time interval (7:05 a.m. to 7:10 a.m.). This process will continue to the destination point. Table 5 shows the results of this calculation. Travel times that are selected are shown in bold font. The final travel time prediction at 7:00 a.m. is 16.56 min for this route. For this case, if we use the travel time estimation at 7:00 a.m., then the result is 14.15 min.

Table 5. Dynamic Travel Time Prediction Results for a Sample Route

TMC Code	Travel Time Estimation at 7:00 a.m.	Travel Time Prediction				Aggregated Travel Time Prediction from the Origin
		7:00–7:05 a.m.	7:05–7:10 a.m.	7:10–7:15 a.m.	7:15–7:20 a.m.	
107+04186	0.31	**0.38**	0.35	0.41	0.35	0.38
107P04186	0.28	**0.29**	0.26	0.29	0.32	0.67
107+04187	0.07	**0.07**	0.06	0.07	0.07	0.74
107P04187	0.49	**0.57**	0.55	0.57	0.52	1.31
107+04188	0.02	**0.01**	0.01	0.01	0.01	1.32
107P04188	0.42	**0.38**	0.36	0.37	0.39	1.70
107+04189	0.05	**0.04**	0.04	0.05	0.05	1.74
107P04189	0.17	**0.16**	0.14	0.16	0.15	1.90
107+04190	0.60	**0.60**	0.54	0.55	0.54	2.50
107P04190	0.10	**0.10**	0.09	0.09	0.09	2.60
107+04191	0.24	**0.27**	0.23	0.23	0.21	2.87
107P04191	0.20	**0.24**	0.19	0.20	0.17	3.11
107+04192	0.11	**0.12**	0.12	0.11	0.09	3.23
107P04192	0.24	**0.30**	0.28	0.26	0.22	3.53
107+04193	0.31	**0.36**	0.38	0.48	0.34	3.89
107P04193	0.74	**0.82**	0.96	1.30	1.08	4.71
107+04194	0.91	**0.75**	0.91	1.13	0.79	**5.46** > 5
107P04194	0.92	0.86	**1.29**	0.91	0.70	6.75
107+04195	1.69	2.02	**2.75**	1.89	1.68	9.50
107P04195	0.85	1.31	**0.93**	0.96	0.72	**10.43** > 10
107+04196	0.09	0.14	0.10	**0.12**	0.12	10.55
107P04196	0.92	1.74	1.42	**1.42**	1.42	11.97
107+04197	0.70	1.88	0.86	**1.03**	0.56	13.00
107P04197	2.84	1.89	3.15	**2.18**	4.73	**15.18** > 15
107+04198	0.86	0.19	0.68	0.68	**1.37**	**16.55**

The dynamic approach is applied to data for the westbound Eisenhower Expressway to predict travel time for a corridor. Figure 29 compares the results of the dynamic and snapshot approaches to predict and calculate travel time for the 25 links of the sample route. Using the dynamic approach could achieve higher accuracy and points are much closer to the 45-degree line. The accuracy of predicting travel time for this corridor is 97.4% and 94.3% for the dynamic and snapshot approaches, respectively.

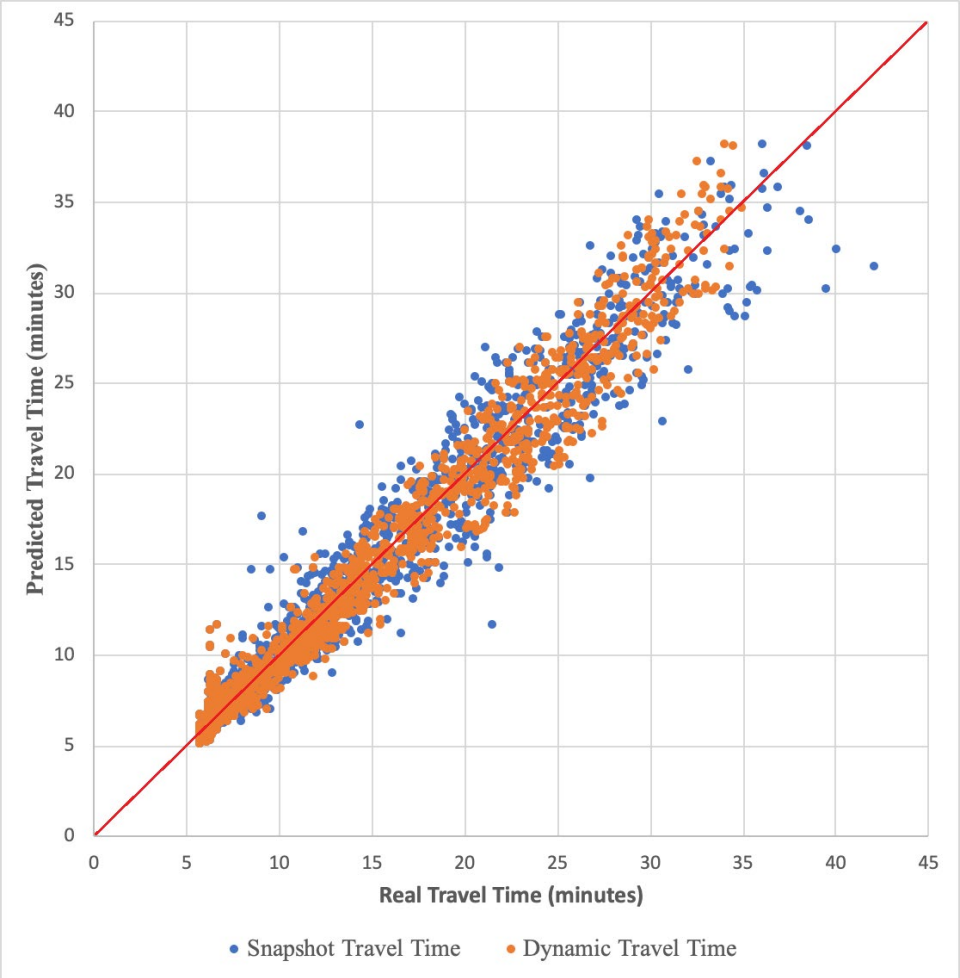


Figure 29. Travel time prediction results for snapshot and dynamic approaches.

FINAL RESULTS FOR ALL HIGHWAYS

All data processing and modeling steps are applied to all highways using available data sources. The studied highways include Eisenhower (W & E), Kennedy 1 (W & E), Kennedy 2 (W & E), Edens (N & S), I-55 (N & S), I-57 (N & S), Bishop (W & E), I-80 (W & E), Dan Ryan 1 (N & S), Dan Ryan 2 (N & S), and Dan Ryan 3 (N & S). For each direction and highway, a model is developed to predict travel time dynamically. For some highways such as the Kennedy and Dan Ryan, more than one model is developed for each direction, because characteristics of the highways are different before and after specific points in which they connect with other highways. Figure 30 shows the location of the studied highways.

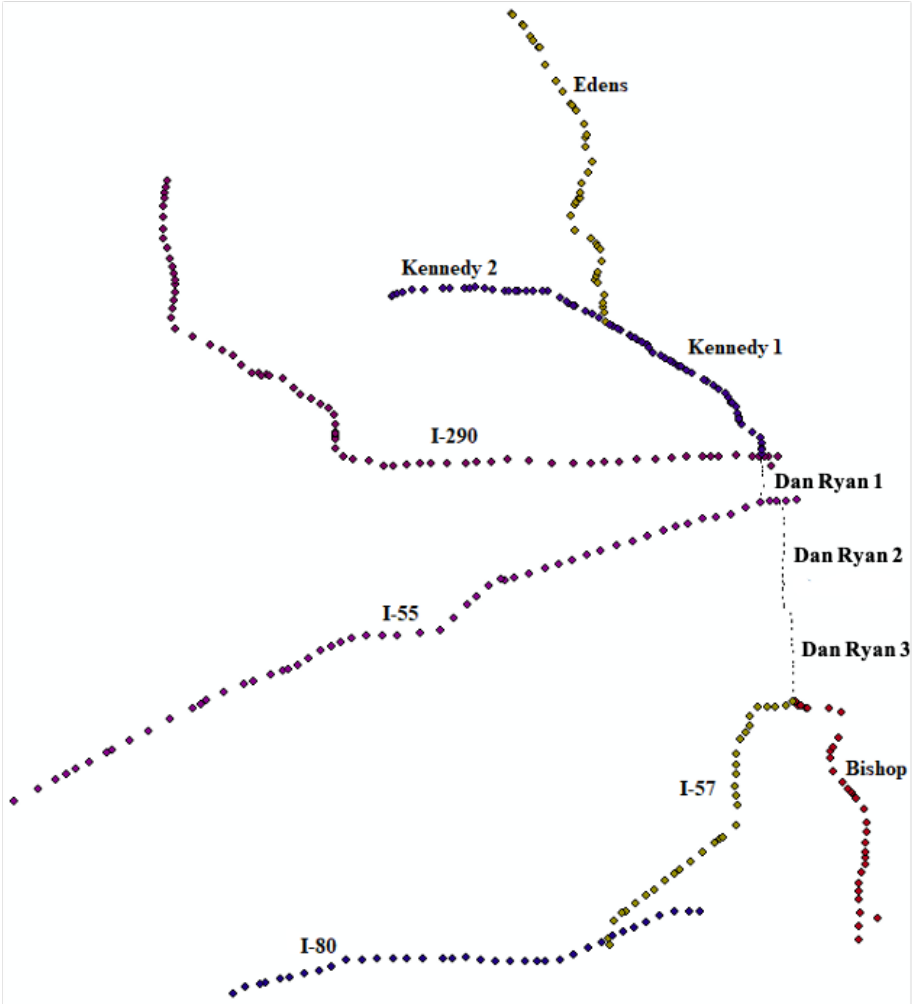


Figure 30. Location of studied northeast Illinois highways.

Table 6 displays the average prediction accuracy of all links in the studied highways. Prediction accuracy of most highways is in an acceptable range. However, prediction accuracy is relatively less in some highways, namely I-57, southbound of Dan Ryan 3, and westbound of I-80. In these highways, the analysis shows that unexpected results are caused by unreliable data sources, especially loop-detector data. This is due to malfunctioning loop detectors and the uncertainty regarding the location of loop detectors, which could cause significant errors. Therefore, the results for these highways are not reliable and need to be further analyzed and studied.

Note that the accuracies reported in Table 6 are the average of instantaneous prediction accuracies for the links of highways. However, in the end, the proposed dynamic approach will be used in practice to predict the travel time for defined corridors. This part significantly improves prediction accuracy of highways' travel time. Each corridor could be defined based on the location of DMS, or any other needs. A sample corridor that has 25 links is investigated previously for the westbound Eisenhower Expressway. Based on Table 6, link-based accuracy for the westbound Eisenhower Expressway is 92%. However, prediction accuracy for the same corridor using the dynamic approach is 97.4%, so using the dynamic approach improves accuracy.

Table 6. Prediction Accuracy of Models for All Highways

Highway	Direction	Average Model Accuracy for All Links (%)
Eisenhower	W	92.0
	E	90.8
Kennedy 1	W	94.1
	E	92.4
Kennedy 2	W	92.6
	E	88.9
Edens	N	80.1
	S	89.9
I-55	N	77.3
	S	85.4
I-57	N	62.3
	S	56.4
Bishop	W	83.1
	E	76.8
I-80	W	63.3
	E	88.5
Dan Ryan 1	N	89.5
	S	90.6
Dan Ryan 2	N	75.9
	S	77.3
Dan Ryan 3	N	90.7
	S	53.8

CHAPTER 6: SUMMARY AND CONCLUSIONS

This study combined multiple data sources, including loop detectors, probe vehicles, weather condition, geometry, roadway incidents, roadwork, and special events. In addition, machine learning methods such as KNN and random forest are employed for short-term prediction of travel time for highways. Accordingly, a model is trained for each technique to predict travel time 5 min, 10 min, and 60 min ahead. A comparison of techniques showed that applying the random forest model for 15 min or shorter prediction horizons is more reasonable, although the prediction accuracy of longer prediction horizons is still acceptable.

Regarding dynamic prediction of travel time, we proposed an algorithm in which travel time of a corridor is calculated by adding up predicted travel time of each link of the corridor. Travel time of a link is predicted for the moment that we expect a vehicle that has entered the corridor to arrive at this link (using aggregated predicted travel time of previous links). The proposed approach is tested and evaluated on highways and showed a significant improvement in the predicted travel time in comparison to the snapshot approach.

Traffic-related variables, especially occupancy, are found to be effective in short-term travel time prediction using loop-detector data. This suggests that among traffic variables collected by loop detectors, occupancy can capture traffic condition better than others. However, fusion of traffic data from different sources could increase prediction accuracy of the models. Weather condition is an important variable in the studies related to traffic condition. In this project, we utilized weather condition and temperature in our models and found them effective in predicting travel time. Finally, roadway incidents, special events, and roadwork are important variables that can have a considerable impact on travel time prediction.

We also generated, analyzed, and inputted the sun glare variable to the models and found that this variable did not improve accuracy significantly. Our analysis shows that although sun glare occurrence could reduce speed, the main impact of sun glare is captured indirectly by other traffic-related variables used in the model. Sun glares mostly occur during peak hours in which traffic speed is already low and, consequently, have less impact on traffic flow.

FUTURE DIRECTIONS

In this study, we tested different data-driven techniques. The best model is selected to predict the travel time of each direction of all studied highways based on the available sources of data. Handling missing data is an important issue that has a considerable impact on accuracy of models, especially data-driven models. Because the main source of data is collected in this study by loop detectors, which include missing and erroneous records, we applied different approaches for data imputation. As mentioned previously, the prediction accuracy for some highway links are not acceptable. This could be due to malfunctioning loop detectors and uncertainty about the exact location of some loop detectors. Therefore, there are two options that are suggested to achieve more accurate reliable models for travel time prediction in all segments.

First, it is suggested to perform comprehensive maintenance on loop detectors, and a clear and reliable resource for loop-detector locations should be provided. Then, developed models could be recalibrated using the new data. Second, a combination of the current approach with advanced deep-learning techniques could be used to utilize traffic-image data and predict travel time with higher accuracy. To this end, a rich source of traffic-image data collected by traffic cameras and archived by IDOT can be analyzed and used through image-processing and supervised deep-learning approaches. This approach could also cover the missing or erroneous values reported by loop detectors.

REFERENCES

AASHTO. 2001. *Policy on Geometric Design of Highways and Streets*. Washington, DC: American Association of State Highway and Transportation Officials.

Abbott-Jard, Michael, Harpal Shah, and Ashish Bhaskar. 2013. “Empirical Evaluation of Bluetooth and Wifi Scanning for Road Transport.” In *Australasian Transport Research Forum (ATRF)*, 36th, 14.

Adeli, Hojjat. 2001. “Neural Networks in Civil Engineering: 1989–2000.” *Computer-Aided Civil and Infrastructure Engineering* 16 (2): 126–42. https://doi.org/10.1111/0885-9507.00219

Auffray, Benjamin, Rue Maurice Audin, Christopher M. Monsere, and Robert L. Bertini. 2008. “An Empirical Investigation of the Impacts of Sun-Related Glare on Traffic Flow.” In *Proc., 87th Annual Meeting of Transportation Research Board*. Washington, DC: Transportation Research Board.

Ben-Akiva, Moshe, Michel Bierlaire, Didier Burton, Haris N. Koutsopoulos, and Rabi Mishalani. 2001. “Network State Estimation and Prediction for Real-Time Traffic Management.” *Networks and Spatial Economics* 1 (3–4): 293–318. https://doi.org/10.1023/A:1012883811652

Blanco-Muriel, Manuel, Diego C. Alarcón-Padilla, Teodoro López-Moratalla, and Martín Lara-Coira. 2001. “Computing the Solar Vector.” *Solar Energy* 70 (5): 431–41. https://doi.org/10.1016/S0038-092X(00)00156-0

Cambridge Systematics, Inc. 2005. *Traffic Congestion and Reliability: Trends and Advanced Strategies for Congestion Mitigation* (FHWA-HOP-05-064). Washington, DC: Federal Highway Administration.

Castillo, Enrique, María Nogal, José María Menendez, Santos Sanchez-Cambronero, and Pilar Jimenez. 2011. “Stochastic Demand Dynamic Traffic Models Using Generalized Beta-Gaussian Bayesian Networks.” *IEEE Transactions on Intelligent Transportation Systems* 13 (2): 565–81. https://doi.org/10.1109/TITS.2011.2173933

Chen, Chenyi, Jianming Hu, Qiang Meng, and Yi Zhang. 2011. “Short-time Traffic Flow Prediction with ARIMA-GARCH Model.” In *2011 IEEE Intelligent Vehicles Symposium (IV)*, 607–12. IEEE.

Chen, Hao, and Hesham A. Rakha. 2012. “Prediction of Dynamic Freeway Travel Times Based on Vehicle Trajectory Construction.” In *2012 15th International IEEE Conference on Intelligent Transportation Systems*, 576–81. IEEE. https://doi.org/10.1109/ITSC.2012.6338825

Chen, Hao, Hesham A. Rakha, and Catherine C. McGhee. 2013. “Dynamic Travel Time Prediction Using Pattern Recognition.” In *20th World Congress on Intelligent Transportation Systems*. TU Delft.

Chen, Mei, and Steven I. J. Chien. 2001. “Dynamic Freeway Travel-time Prediction with Probe Vehicle Data: Link Based versus Path Based.” *Transportation Research Record* 1768 (1): 157–61.

Chen, Peng, Rui Tong, Guangquan Lu, and Yunpeng Wang. 2018. “Exploring Travel Time Distribution and Variability Patterns Using Probe Vehicle Data: Case Study in Beijing.” *Journal of Advanced Transportation*. https://doi.org/10.1155/2018/3747632

Chien, Steven I-Jy, and Chandra Mouly Kuchipudi. 2003. “Dynamic Travel Time Prediction with Real-Time and Historic Data.” *Journal of Transportation Engineering* 129 (6): 608–16.

Churchill, Andrew M., Yorghos Tripodis, and David J. Lovell. 2012. “Sun Glare Impacts on Freeway

Congestion: Geometric Model and Empirical Analysis." *Journal of Transportation Engineering* 138 (10): 1196–1204. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000418

Cline, David, Henry W. Hofstetter, and John R. Griffin. 1997. *Dictionary of Visual Science*. Oxford, UK: Butterworth-Heinemann.

Cremer, Michael. 1995. "On the Calculation of Individual Travel Times by Macroscopic Models." In *Pacific Rim TransTech Conference. 1995 Vehicle Navigation and Information Systems Conference Proceedings. 6th International VNIS. A Ride into the Future*, 187–93. IEEE.

Dharia, Abhijit, and Hojjat Adeli. 2003. "Neural Network Model for Rapid Forecasting of Freeway Link Travel Time." *Engineering Applications of Artificial Intelligence* 16 (7–8): 607–13. https://doi.org/10.1016/j.engappai.2003.09.011

Du, Lili, Srinivas Peeta, and Yong Hoon Kim. 2012. "An Adaptive Information Fusion Model to Predict the Short-Term Link Travel Time Distribution in Dynamic Traffic Networks." *Transportation Research Part B: Methodological* 46 (1): 235–52. https://doi.org/10.1016/j.trb.2011.09.008

Edara, Praveen, Roozbeh Rahmani, Henry Brown, and Carlos Sun. 2017. "Traffic Impact Assessment of Moving Work Zone Operations." Smart Work Zone Deployment Initiative.

El Faouzi, Nour-Eddin, Lawrence A. Klein, and Olivier De Mouzon. 2009. "Improving Travel Time Estimates from Inductive Loop and Toll Collection Data with Dempster–Shafer Data Fusion." *Transportation Research Record* 2129 (1): 73–80. https://doi.org/10.3141%2F2129-09

Fan, Shu-Kai S., Chuan-Jun Su, Han-Tang Nien, Pei-Fang Tsai, and Chen-Yang Cheng. 2018. "Using Machine Learning and Big Data Approaches to Predict Travel Time Based on Historical and Real-Time Data from Taiwan Electronic Toll Collection." *Soft Computing* 22 (17): 5707–18. https://doi.org/10.1007/s00500-017-2610-y

Giesen, Juergen. 2018. "'JavaScript Sun Calculator.'" Last modified January 4, 2018. http://www.jgiesen .de/astro/astroJS/riseset/index.htm.

Goodwin, Lynette C. 2002. "Weather Impacts on Arterial Traffic Flow." *Mitretek Systems Inc*, 4–8.

Han, Jiawei, Jian Pei, and Micheline Kamber. 2011. *Data Mining: Concepts and Techniques*. Waltham, MA: Morgan Kaufmann.

Huisken, Giovanni, and Eric C. van Berkum. 2003. "A Comparative Analysis of Short-Range Travel Time Prediction Methods." In *82nd Annual Meeting of the Transportation Research Board*, Washington, DC.

Hussein, Haitham M., Morgan M. Brown, and Amanda A. Herrmann. 2019. "Abstract WP325: A Travel Time and Efficiency Simulation for Patient Transport to Primary and Comprehensive Stroke Centers." *Stroke* 50 (Suppl_1): AWP325–AWP325.

INRIX. 2019. "Congestion Costs Each American 97 hours, $1,348 A Year." INRIX, February 11, 2019. https://inrix.com/press-releases/scorecard-2018-us/

Jurado-Piña, R., J. M. Pardillo-Mayora, and R. Jiménez. 2010. "Methodology to Analyze Sun Glare Related Safety Problems at Highway Tunnel Exits." *Journal of Transportation Engineering* 136 (6): 545–53. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000113

Kwon, Jaimyoung, Benjamin Coifman, and Peter Bickel. 2000. “Day-to-Day Travel-Time Trends and Travel-Time Prediction from Loop-Detector Data.” *Transportation Research Record* 1717 (1): 120–29. https://doi.org/10.3141/1717-15

Leshem, Guy, and Yaacov Ritov. 2007. “Traffic Flow Prediction Using Adaboost Algorithm with Random Forests as a Weak Learner.” In *Proceedings of World Academy of Science, Engineering and Technology* 19: 193–98. Citeseer.

Li, Xiaojiang, Bill Yang Cai, Waishan Qiu, Jinhua Zhao, and Carlo Ratti. 2019. “A Novel Method for Predicting and Mapping the Occurrence of Sun Glare Using Google Street View.” *Transportation Research Part C: Emerging Technologies* 106: 132–44.

Michalsky, Joseph J. 1988. “The Astronomical Almanac’s Algorithm for Approximate Solar Position (1950–2050).” *Solar Energy* 40 (3): 227–35. https://doi.org/10.1016/0038-092X(88)90045-X

Mori, Usue, Alexander Mendiburu, Maite Álvarez, and Jose A. Lozano. 2015. “A Review of Travel Time Estimation and Forecasting for Advanced Traveller Information Systems.” *Transportmetrica A: Transport Science* 11 (2): 119–57. https://doi.org/10.1080/23249935.2014.932469

Nikovski, D., N. Nishiuma, Y. Goto, and H. Kumazawa. 2005. “Univariate Short-Term Prediction of Road Travel Times.” In *Proceedings. 2005 IEEE Intelligent Transportation Systems*, 1074–79. IEEE.

NOAA. n.d. “NOAA Solar Calculator.” Accessed May 16, 2020. https://www.esrl.noaa.gov/gmd/grad/solcalc/index.html.

Nookala, Lalit Sivanandan. 2006. “Weather Impact on Traffic Conditions and Travel Time Prediction.” Citeseer.

Papageorgiou, Markos, Ioannis Papamichail, Albert Messmer, and Yibing Wang. 2010. “Traffic Simulation with METANET.” In *Fundamentals of Traffic Simulation*, edited by Jaume Barceló, 399–430. Springer. https://doi.org/10.1007/978-1-4419-6142-6_11

Qiao, Wenxin, Ali Haghani, and Masoud Hamedi. 2012. “Short-Term Travel Time Prediction Considering the Effects of Weather.” *Transportation Research Record* 2308 (1): 61–72. https://doi.org/10.3141%2F2308-07

———. 2013. “A Nonparametric Model for Short-Term Travel Time Prediction Using Bluetooth Data.” *Journal of Intelligent Transportation Systems* 17 (2): 165–75. https://doi.org/10.1080/15472450.2012.748555

Reda, Ibrahim, and Afshin Andreas. 2004. “Solar Position Algorithm for Solar Radiation Applications.” *Solar Energy* 76 (5): 577–89. https://doi.org/10.1016/j.solener.2003.12.003

Sánchez-Cambronero, Santos, Pilar Jiménez, Ana Rivas, and Inmaculada Gallego. 2017. “Plate Scanning Tools to Obtain Travel Times in Traffic Networks.” *Journal of Intelligent Transportation Systems* 21 (5): 390–408. https://doi.org/10.1080/15472450.2017.1298037

Schmitt, Erick J., and Hossein Jula. 2007. “On the Limitations of Linear Models in Predicting Travel Times.” In *2007 IEEE Intelligent Transportation Systems Conference*, 830–35. IEEE.

Shen, Luou. 2008. “Freeway Travel Time Estimation and Prediction Using Dynamic Neural Networks.” Florida International University.

Shen, Luou, and Min Huang. 2011. “Assessing Dynamic Neural Networks for Travel Time Prediction.” In *International Conference on Applied Informatics and Communication*, 469–77. Springer. https://doi.org/10.1007/978-3-642-23214-5_62

Shepard, Frank D. 1996. *Reduced Visibility Due to Fog on the Highway*. Vol. 228. Transportation Research Board.

Simroth, Axel, and Henryk Zahle. 2010. “Travel Time Prediction Using Floating Car Data Applied to Logistics Planning.” *IEEE Transactions on Intelligent Transportation Systems* 12 (1): 243–53. https://doi.org/10.1109/TITS.2010.2090521

Soriguera, Francesc, and Francesc Robusté. 2011. “Estimation of Traffic Stream Space Mean Speed from Time Aggregations of Double Loop Detector Data.” *Transportation Research Part C: Emerging Technologies* 19 (1): 115–29.

Soriguera, Francesc, and Francesc Robusté. 2010. “Requiem for Freeway Travel Time Estimation Methods Based on Blind Speed Interpolations between Point Measurements.” *IEEE Transactions on Intelligent Transportation Systems* 12 (1): 291–97. https://doi.org/10.1109/TITS.2010.2095007

Taylor, Nicholas B. 2003. “The CONTRAM Dynamic Traffic Assignment Model.” *Networks and Spatial Economics* 3 (3): 297–322. https://doi.org/10.1023/A:1025394201651

SunEarthTools. n.d. “Tools for Consumers and Designers of Solar.” Accessed May 16, 2020. https://www.sunearthtools.com/dp/tools/pos_sun.php?lang=en#top.

van Hinsbergen, C. P. I. J., J. W. C. van Lint, and H. J. van Zuylen. 2009. “Bayesian Committee of Neural Networks to Predict Travel Times with Confidence Intervals.” *Transportation Research Part C: Emerging Technologies* 17 (5): 498–509. https://doi.org/10.1016/j.trc.2009.04.007

van Lint, J. W. C. 2004. *Reliable Travel Time Prediction for Freeways*. Netherlands TRAIL Research School.

van Lint, J. W. C., and C. P. I. J. Van Hinsbergen. 2012. “Short-Term Traffic and Travel Time Prediction Models.” *Artificial Intelligence Applications to Critical Transportation Issues* 22 (1): 22–41.

van Lint, J. W. C., S. P. Hoogendoorn, and Henk J. van Zuylen. 2002. “Freeway Travel Time Prediction with State-Space Neural Networks: Modeling State-Space Dynamics with Recurrent Neural Networks.” *Transportation Research Record* 1811 (1): 30–39. https://doi.org/10.3141%2F1811-04

Wang, Min, and Qing Ma. 2014. “Dynamic Prediction Method of Route Travel Time Based on Interval Velocity Measurement System.” In *Proceedings of 2014 IEEE International Conference on Service Operations and Logistics, and Informatics*, 172–76. IEEE. https://doi.org/10.1109/SOLI.2014.6960714

Wu, Chun-Hsin, Jan-Ming Ho, and Der-Tsai Lee. 2004. “Travel-Time Prediction with Support Vector Regression.” *IEEE Transactions on Intelligent Transportation Systems* 5 (4): 276–81.

Wunderlich, Karl E., David E. Kaufman, and Robert L. Smith. 2000. “Link Travel Time Prediction for Decentralized Route Guidance Architectures.” *IEEE Transactions on Intelligent Transportation Systems* 1 (1): 4–14. https://doi.org/10.1109/6979.869017

Xia, Jingxin, Mei Chen, and Zhendong Qian. 2010. “Predicting Freeway Travel Time under Incident

Conditions." *Transportation Research Record* 2178 (1): 58–66. https://doi.org/10.3141%2F2178-07

Yang, Jiann-Shiou. 2005a. "A Study of Travel Time Modeling via Time Series Analysis." In *Proceedings of 2005 IEEE Conference on Control Applications, 2005. CCA*, 855–60. IEEE.

———. 2005b. "Travel Time Prediction Using the GPS Test Vehicle and Kalman Filtering Techniques." In *Proceedings of the 2005, American Control Conference,* 2128–33. IEEE.

Yildirimoglu, Mehmet, and Nikolas Geroliminis. 2013. "Experienced Travel Time Prediction for Congested Freeways." *Transportation Research Part B: Methodological* 53: 45–63. https://doi.org/10.1016/j.trb.2013.03.006

Yildirimoglu, Mehmet, and Kaan Ozbay. 2012. "Comparative Evaluation of Probe-Based Travel Time Prediction Techniques under Varying Traffic Conditions." *Transportation Research Board 91st Annual Meeting*. Washington DC.

Yu, Bin, Xiaolin Song, Feng Guan, Zhiming Yang, and Baozhen Yao. 2016. "K-Nearest Neighbor Model for Multiple-Time-Step Prediction of Short-Term Traffic Condition." *Journal of Transportation Engineering* 142 (6): 4016018. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000816

Yu, Young Jung, and Mi-Gyung Cho. 2008. "A Short-Term Prediction Model for Forecasting Traffic Information Using Bayesian Network." In *2008 Third International Conference on Convergence and Hybrid Information Technology* 1: 242–47. IEEE.

Yuan, Dian, Ardeshir Faghri, and Katherine Partridge. 2019. "A Study on Applications and Case Studies Regarding Bluetooth Technology for Travel Time Measurement." *Journal of Transportation Technologies* 10 (1): 65–87.

Zeng, Xiaosi. 2011. "Dynamically Predicting Corridor Travel Time Under Incident Conditions Using a Neural Network Approach." Texas A&M University.

Zhang, Xiaoyan, and John A. Rice. 2003. "Short-Term Travel Time Prediction." *Transportation Research Part C: Emerging Technologies* 11 (3–4): 187–210. https://doi.org/10.1016/S0968-090X(03)00026-3

Zhao, Jiandong, Yuan Gao, Jinjin Tang, Lingxi Zhu, and Jiaqi Ma. 2018. "Highway Travel Time Prediction Using Sparse Tensor Completion Tactics and Nearest Neighbor Pattern Matching Method." *Journal of Advanced Transportation*. https://doi.org/10.1155/2018/5721058

Zheng, Fangfang, and Henk Van Zuylen. 2013. "Urban Link Travel Time Estimation Based on Sparse Probe Vehicle Data." *Transportation Research Part C: Emerging Technologies* 31: 145–57. https://doi.org/10.1155/2016/7348705


Illinois Department
of Transportation


ILLINOIS